

Semantics Disentangling for Cross-Modal Retrieval

Zheng Wang¹, Xing Xu¹, *Member, IEEE*, Jiwei Wei¹, Ning Xie¹, Yang Yang¹, *Senior Member, IEEE*,
and Heng Tao Shen², *Fellow, IEEE*

Abstract—Cross-modal retrieval (e.g., query a given image to obtain a semantically similar sentence, and vice versa) is an important but challenging task, as the heterogeneous gap and inconsistent distributions exist between different modalities. The dominant approaches struggle to bridge the heterogeneity by capturing the common representations among heterogeneous data in a constructed subspace which can reflect the semantic closeness. However, insufficient consideration is taken into the fact that learned latent representations are actually heavily entangled with those semantic-unrelated features, which obviously further compounds the challenges of cross-modal retrieval. To alleviate the difficulty, this work makes an assumption that the data are jointly characterized by two independent features: semantic-shared and semantic-unrelated representations. The former presents characteristics of consistent semantics shared by different modalities, while the latter reflects the characteristics with respect to the modality yet unrelated to semantics, such as background, illumination, and other low-level information. Therefore, this paper aims to disentangle the shared semantics from the entangled features, and thus the purer semantic representation can promote the closeness of paired data. Specifically, this paper designs a novel Semantics Disentangling approach for Cross-Modal Retrieval (termed as *SDCMR*) to explicitly decouple the two different features based on variational auto-encoder. Next, the reconstruction is performed by exchanging shared semantics to ensure the learning of semantic consistency. Moreover, a dual adversarial mechanism is designed to disentangle the two independent features via a pushing-and-pulling strategy. Comprehensive experiments on four widely used datasets demonstrate the effectiveness and superiority of the proposed *SDCMR* method by achieving a new bar on performance when compared against 15 state-of-the-art methods.

Index Terms—Cross-modal retrieval, semantics disentangling, dual adversarial mechanism, subspace learning.

I. INTRODUCTION

DUE to the popularization of 5G and smart devices, people are surrounded by the exponential explosion of multimedia data, e.g., image, text, video, audio, etc. Multimedia data that collectively describe a topic can help people to understand the content quickly and thoroughly, which spurs the demand for cross-modal retrieval [1]. Unlike traditional retrieval focusing on the same modality, cross-modal retrieval is a more challenging task as it is impossible to measure the cross-modal similarity directly [2], [3], [4], [5] for the existence of heterogeneous gaps and inconsistent distributions. An intuitive solution for capturing the latent similarity of different modalities is to map the heterogeneous data into a well-constructed subspace through learning the common representation, shown as Fig. 1(a). Then, the semantic similarity among different modalities can be measured by Euclidean distance or cosine similarity. As the optimization goes on, semantically similar cross-modal data will be pulled closer together in the constructed subspace while those dissimilar data will be pushed further apart.

Thanks to its powerful nonlinear learning ability, deep learning [6], [7], [8] has gradually replaced traditional methods [9], [10], [11] to become the main remedy for cross-modal retrieval [12], [13], [14], [15]. Adversarial Cross-Modal Retrieval (ACMR) [16] is a typical work for common subspace learning methods in recent years, which firstly modeled the intra-modal correlation by generating the corresponding representations and then explored the cross-modal correlation. Its following work [17], [18], [19], [20] further improved the mining of those heterogeneous correlations. These previous methods undoubtedly made remarkable progress, but they are actually plagued by the phenomenon that the learned common representations are the entanglement with much semantic-unrelated information. The correlation measurement is accordingly interfered with to degrade the performance of cross-modal retrieval. The semantic-unrelated information can be roughly divided into two categories: shape, height, and other appearances associated with the modality; illumination, background, and other low-level features. Evidently, the undecoupled information is not conducive to shortening the heterogeneous gap.

To better clarify the motivation, this paper takes text-to-image retrieval as an example to further elucidate the performance degradation caused by the entanglement. As shown

Manuscript received 22 August 2022; revised 28 May 2023 and 25 September 2023; accepted 5 February 2024. Date of publication 18 March 2024; date of current version 25 March 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 62306065, Grant 62220106008, and Grant 62020106008; in part by the Sichuan Science and Technology Program, China, under Grant 2023YFG0289; and in part by the Guangdong Basic and Applied Basic Research Foundation under Grant 2022A1515110576 and Grant 2023A1515140104. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Wangmeng Zuo. (*Corresponding author: Yang Yang.*)

Zheng Wang and Yang Yang are with the Center for Future Multimedia and the School of Computer Science and Engineering, University of Electronic Science and Technology of China (UESTC), Chengdu 611731, China, and also with the Institute of Electronic and Information Engineering, UESTC in Guangdong, Dongguan 523808, China (e-mail: zh_wang@hotmail.com; dlyyang@gmail.com).

Xing Xu, Jiwei Wei, and Ning Xie are with the Center for Future Multimedia and the School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China (e-mail: xing.xu@uestc.edu.cn; mathematic6@gmail.com; seanxieming@gmail.com).

Heng Tao Shen is with the Center for Future Multimedia and the School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China, and also with the Peng Cheng Laboratory, Shenzhen 518066, China (e-mail: shenhengtao@hotmail.com).

Digital Object Identifier 10.1109/TIP.2024.3374111

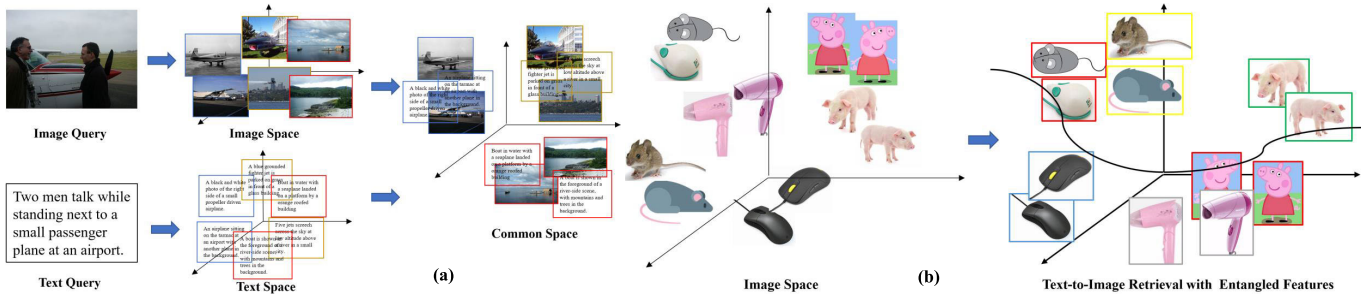


Fig. 1. (a) Flow diagram of common subspace learning; (b) Illumination of entanglement with semantic-unrelated information.

in Fig.1(b), the nose of Peppa pig is shaped like a hair dryer and the similar case also occurs between the rat and the photoelectric mouse. However, existing methods did not separate the consistently and irrelevantly semantic features and thus the entwined representations will lead to the instance distorted in the feature manifold, which causes incorrect matching of similarity. For example, a textual query about a rat may yield some optical mice, since the appearance interferes with the learning of common representation. Similarly, several images of Peppa pig can be obtained when querying with a sentence about a hair dryer, as the nose's shape is entangled with semantics of the hair drier. In consequence, disentangling the semantics from those coupled features is another challenge for cross-modal retrieval.

To tackle the aforementioned challenge, this work makes an assumption that the representation of each modality is highly entangled by two independent features: semantic-shared and semantic-unrelated representations. Particularly, the former presents consistent semantic features shared by different modalities, while the latter reflects the characteristics with respect to the modality yet unrelated to semantics, such as appearance and low-level information. Therefore, this paper aims to disentangle the shared semantics from the entangled features, which thus helps the representations to preserve semantic consistency and promote the closeness of paired data. Specifically, this paper formulates a novel **Semantics Disentangling approach for Cross-Modal Retrieval** (abbr. as **SDCMR**) to explicitly decouple the two different features. Driven by the decoupling of explanatory factors in representation learning [21], [22], the SDCMR method utilizes M variational auto-encoders (VAE) to generate the semantic-unrelated information with respect to the modality, where M means the category number of modalities. Meanwhile, an additional VAE network shared by all modalities is designed to extract the latent common semantic features. Subsequently, a dual-adversarial mechanism utilizes a pushing-and-pulling strategy to extract semantic-consistent information more purely, while others is pushed into semantically unrelated features. This adversarial process also promotes the disentanglement of the entangled representations, thus boosting the performance of retrieval only with the semantic-consistent features. In the end, the M VAEs refactor the disentangled semantic-unrelated features and the exchanged semantic-shared features across modalities into the original representations, which can encourage a better reconstruction.

The differences with the similar work P3S (Private-Shared Subspaces Separation) [23] are the following: 1) P3S constructs its private and shared encoders through several fully connected layers, while our SDCMR approach is first to introduce the theory of disentangling to learn two independent representations by constructing multi-variable VAE networks. In contrast, our semantics disentangling is more interpretable. 2) Additionally, we propose a novel dual adversarial mechanism, which applies a push-and-pull strategy to squeeze out the semantic-shared representation from the semantic-unrelated feature, while expelling semantically independent information out of the semantic consistency to promote the decoupling.

Our main contributions can be summarized as follows:

- We introduce the theory of decoupling into cross-modal retrieval. The disentangled representation can bypass the interference of non-semantic information and thus better measure the semantic similarity.
- A novel semantic representation disentangling architecture called SDCMR is designed, which contains two types of associate subnetworks: M variational auto-encoders that learn the latent semantic-unrelated features, and an extra VAE shared by all modalities for the semantic-shared features learning.
- A dual adversarial mechanism is proposed to effectively separate the two representations from the entangled features, which boosts the cross-modal correlation of the heterogeneous data only with the semantic-shared features.

II. RELATED WORK

Existing methods are discussed from four aspects.

A. Traditional Approaches

The previous strategies for common representation learning were mainly through searching linear projections [24], [25], [26]. The most representative work is canonical correlation analysis (CCA) [27], which maximized the similarity across heterogeneous data by mapping them into a low-dimensional common subspace. Subsequently, a series of extensions based on CCA [27] were formulated to deal with cross-modal retrieval. For example, semantic labels were combined with CCA to capture the correlations of different modalities [9], [28]. Multi-view CCAs [25], [29] were developed to keep highly consistent with the most canonical variables

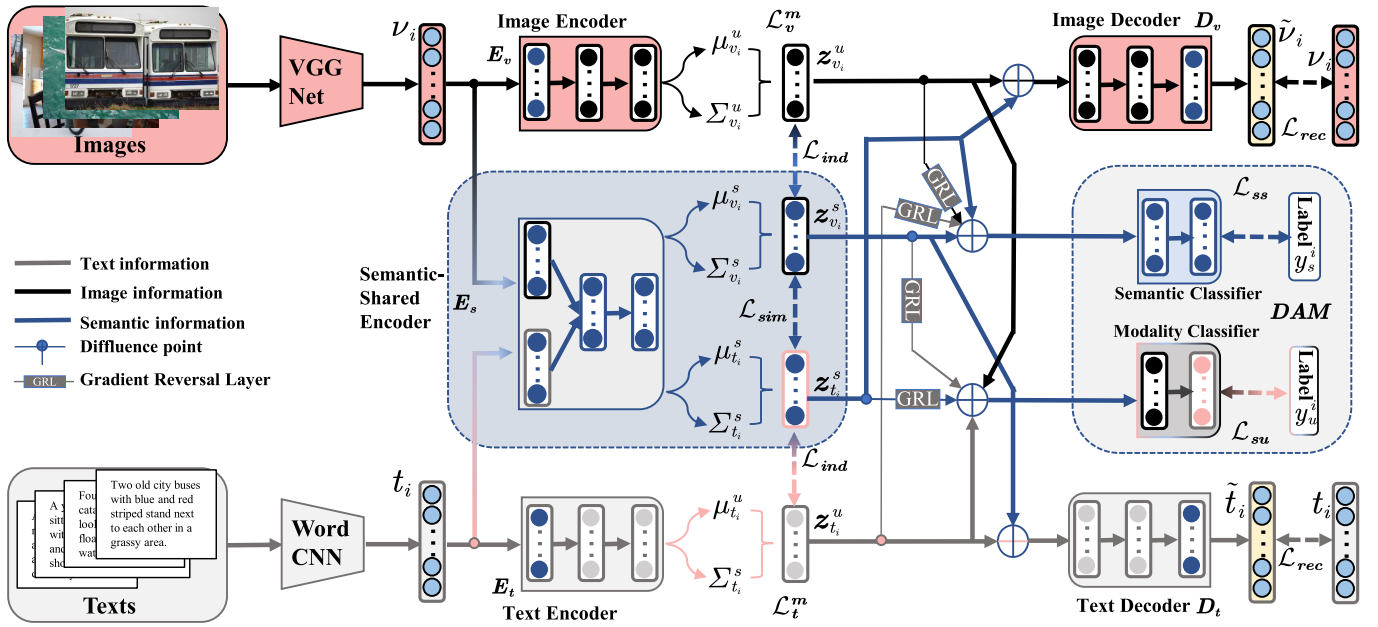


Fig. 2. The framework of our SDCMR for image and text contains four parts: two specific encoders E_v , E_t for the semantic-unrelated representation learning; a semantic-shared encoder E_s for the semantic-shared representation extracting; a dual adversarial mechanism **DAM** to pull the semantic relevant information from the entangled features into semantic-shared representation and yet to expel the irrelevant information into semantic-unrelated representations; two decoders D_v , D_t for the input reconstruction with swapped semantic-shared representations. Best viewed with colors and thickness of the flows.

from each view of the data. Besides, several advanced strategies including cross-modal factor analysis [10], half-quadratic optimization for feature selection [30], coupled dictionary learning for supervised sparse coding [24], were in succession introduced to minimize the distances of the pairwise data.

However, their common drawback is that it takes intensive labor to pair cross-modal data. To cope with this difficulty, several methods [11], [31] directly utilized unlabeled images or texts crawled on the web to build partially labeled graphs, thus improving cross-modal correlation, albeit not significantly compared to deep learning methods.

B. DNN-Based Approaches

In recent decades, deep learning has made great progress in cross-modal retrieval due to its strong ability for nonlinear representation learning [32]. Specifically, the correspondence auto-encoder (Corr-AE) [12] first incorporated the structure of auto-encoder and associations cost to build a two-way network, which effectively facilitated the grasp of correlation for pairwise data. Cross-media Multiple Deep Networks (CMDN) [33] first exploited the complex cross-media correlation with hierarchical learning. Then its extension CCL [34] fused the coarse-grained instances and fine-grained patches to make cross-modal correlation more precise. The concurrent work, cross-modal hybrid transfer network (CHTN) [35], introduced two sub-networks for knowledge transferring and semantic preserving between cross-modal data, respectively. In addition, intra-class low-rank constraint [36] and online similarity learning framework [37], [38] were designed to better solve the problem of intrinsic heterogeneity and information imbalance. Differently, deep supervised cross-modal retrieval (DSCMR) [13] simply utilized several fully

connection layers shared by different modalities yet effectively captured the cross-modal correlation.

Nevertheless, they ignored the interference of semantically irrelevant features. That is, the learned common representation is highly coupled with modality-dependent features, including appearances, shapes, low-level features, and so on, which thus largely limited the performance of cross-modal retrieval.

C. GAN-Based Approaches

Recently, benefiting from the strong capability of generation and data distribution fitting, the generative adversarial network (GAN) [39] has been broadly used in various tasks of computer vision, such as image generation [40] and style transferring [41]. Adversarial cross-modal retrieval (ACMR) [16] pioneered adversarial learning with an interplay of a min-max game to cross-modal retrieval, which can effectively bridge the heterogeneity gap. Then various adversarial learning strategies for cross-modal retrieval, including cross-modal generative adversarial framework (CM-GAN) [17], modal-adversarial hybrid transfer network (MHTN) [18], Ternary adversarial networks with self-supervision (TANSS) [20], Joint Feature Synthesis and Embedding (JFSE) [19] and category alignment adversarial learning (CAAL) [42], emerged continuously.

Analyzed as [19], the vanilla GAN models encounter some difficulties, such as unstable training, meaningless synthetic features, and lack of extensibility. To this end, more complex GAN-based methods are required, which yet limits their practicality even with their excellent performance.

D. Semantics Disentangling

Although previous methods [13], [18], [23], [37], [42] work remarkably well, they actually suffer from the entanglement

with semantic-unrelated features. On the contrary, semantics disentangling methods take the interference of those irrelevant information into consideration, and aim to explicitly split the individuality and commonality of heterogeneous data. Lately, feature disentangling attracts extensive attention in the field of computer vision, such as image classification [43], person re-identification [44], single image de-raining [45] and zero-shot learning [46], and also proliferates into natural language understanding [47]. Driven by the benefit of disentangling methods, several works transferred them into the field of cross-modal retrieval to bridge vision and language. Particularly, the MS2GAN method [48] was presented to learn the specific and shared features for each modality with two sub-networks. The representation separation adversarial network (ReSAN) [49] separated the initial representations into common and private parts via a single adversarial training mechanism. The similar work P3S [23] eliminated the elementary interference by learning a shared subspace and a private subspace.

Despite their significant progress on the correlation capture between heterogeneous data, the semantics disentangling strategies need to be further explored to thoroughly decouple the semantics from entangled features in cross-modal retrieval. Differently, the proposed SDCMR method combines the interpretable factor decomposition and the theory of VAE to disentangle the semantics through a novel dual adversarial mechanism, which can filter out the interference of non-semantic features.

III. THE PROPOSED SDCMR APPROACH

A. Problem Statement and Notations

The core idea is illustrated by taking bimodal data of images and text as an example. Particularly, the pairwise training split is formulated as $\mathcal{O}_{tr} = \{o_i\}_{i=1}^N$ and $o_i = (v_i, t_i)$, where $v_i \in \mathbb{R}^{d_v}$ is the d_v -dimensional feature of i -th image, and $t_i \in \mathbb{R}^{d_t}$ means that of its description. A semantic label $y_i \in \mathcal{Y} = \{y_i\}_{i=1}^C$ with C different labels is allocated to the o_i instance. Beyond that, a test set \mathcal{O}_{te} with N' instances per modality is also prepared well, viz. $\{v'_j\}_{j=1}^{N'}$ and $\{t'_j\}_{j=1}^{N'}$.

The goal of this work is aimed to seek an effective disentangling strategy for the semantic consistency among different modalities, since the initial representations obtained from the off-the-shelf backbones, such as VGG-19 [50] and WordCNN [51], are highly entangled with much non-semantic information. With this disentanglement, we can more accurately and effectively learn semantic representations shared by heterogeneous data to minimize the interference caused by non-semantic information dependent on modality. At the inference, we employ the well-trained SDCMR to uncouple \mathcal{O}_{te} for better cross-modal retrieval.

B. Model Architecture

The overall framework of the proposed SDCMR approach is depicted in Fig. 2 with four core parts. The first part contains M specific encoders for all modalities to learn the semantic-unrelated representation. M is set to 2 since we take bimodal data (image and text) to illustrate our method. Namely, the two encoders E_v, E_t are for image and text respectively.

The second part is a semantic-shared encoder E_s to extract the semantic consistent representation across modalities. Note that the three encoders are all based on the theory of Variational Auto-Encoder (VAE), which is discussed below. Another key component is a dual adversarial mechanism **DAM**, which promotes the disentangling via a pulling-and-pushing strategy. The instance reconstruction is essential to ensure the correctness of the decoupling, which mainly includes two decoders D_v, D_t for image and text respectively.

Motivated by representation disentangling [21], [22], we also consider that the features of any modality are highly entangled by two independent latent variables: modality-dependent variable z^u for unrelated semantic information and modality-agnostic variable z^s for semantic consistent information shared by all modalities. Thus, we attempt to separate the semantic representation \mathbf{z}^s formulated by variable z^s from the non-semantic representation \mathbf{z}^u constructed by z^u . Specifically, we modify the structure of VAE to equip each modality with two encoders and one decoder, which are all made up of several fcs (fully connected layers). Take the image branch as an example in Fig. 2, E_v and E_s constitute the encoder of our modified VAE framework. The modality-dependent encoders $E_*, * = v, t$ aim to extract the non-semantic representations \mathbf{z}_*^u from v_i and t_i , such as illumination, background, irrelevant objects, noisy words and others. At the same time, the semantic encoder E_s captures the cross-modal association by generating the latent semantic representations \mathbf{z}_*^s , which is consistent among heterogeneous data. To better preserve the property of independence in two latent representations, an additional orthogonality constraint is imposed on them, formulated as $\mathbf{z}_*^s \perp \mathbf{z}_*^u$. At the reconstruction, the semantic representations are swapped between image and text to further maintain semantic consistency. In other words, $\mathbf{z}_v = \{\mathbf{z}_v^u, \mathbf{z}_v^s\}$, $\mathbf{z}_t = \{\mathbf{z}_t^u, \mathbf{z}_t^s\}$ are feed into the decoder D_* to robustly reconstruct the original coupled features. Finally, our dual adversarial mechanism **DAM** can pull the semantic relevant information from the entangled features into semantic-shared representation \mathbf{z}^s and yet expel the irrelevant information into semantic-unrelated representations \mathbf{z}^u . This pull-and-push strategy can promote the classification of \mathbf{z}_*^s and \mathbf{z}_*^u into right labels. Next, we will go into the details of each part.

C. Multi-Latent Variables Reconstruction

In this paper, we employ VAE [52] as the basic framework for each modality, which consists of an encoder for obtaining the low dimensional latent variable z from the input data and a decoder in charge of reconstructing the input from z . Theoretically, the VAE aims to find the true conditional probability distribution $p(z|x)$ by variational inference. However, the closest proxy posterior $q(z|x)$ in fact is utilized to approximate the intractable $p(z|x)$ by minimizing the distance of $q(z|x)$ and $p(z|x)$ [53]. Therefore, the objective function is transformed into a variational lower bound on the marginal likelihood as follows:

$$\mathcal{L}'_{ELBO} = -D_{KL}(q(z|x) \parallel p(z)) + \mathbb{E}_{q(z|x)} [\log p(x|z)], \quad (1)$$

where the former item is for Kullback-Leibler divergence [22] and the latter defines the error of reconstruction. The two conditional probability distributions $p(x|z)$ and $q(z|x)$ are the encoder and decoder for a standard VAE respectively. In addition, $p(z)$ models the prior distribution, following a multivariate Gaussian distribution. For brevity, similar to other VAE literatures [53], [54], [55], our proposed SDCMR method applies a re-parameterization trick, where $\Sigma = \mathbf{I}$ and $\mu = \mathbf{0}$ are respectively the variance and mean for the posterior distribution $q(z|x) \sim \mathcal{N}(\mu, \Sigma)$.

To achieve the goal of semantics disentangling, we further factor the latent variable $z_*, * = v, t$ of each modality into two independent variables z_*^u and z_*^s . Hence, the new objective function for our modified VAE structure can be derived as:

$$\begin{aligned} \mathcal{L}_v^m = & -D_{KL}(q(z_v^u|v) \parallel p(z_v^u)) \\ & -D_{KL}(q(z_v^s|v) \parallel p(z_v^s)) \\ & + \mathbb{E}_{q(z_v^u, z_v^s|v)} [\log p(v|z_v^u, z_v^s)], \end{aligned} \quad (2)$$

where $p(z_v^u)$ and $p(z_v^s)$ are also following the normal distribution as Eq. 1. The re-parameterization trick is also employed for the approximator of q distribution to encode the input features of a given image v_i into z_v^u and z_v^s respectively.

$$\begin{aligned} \mathcal{L}_t^m = & -D_{KL}(q(z_t^u|t) \parallel p(z_t^u)) \\ & -D_{KL}(q(z_t^s|t) \parallel p(z_t^s)) \\ & + \mathbb{E}_{q(z_t^u, z_t^s|t)} [\log p(t|z_t^u, z_t^s)]. \end{aligned} \quad (3)$$

By analogy, we can easily derive the multi-latent variable object function \mathcal{L}_t^m of the modified VAE for the text data, elaborated as Eq. 3.

$$\mathcal{L}_{ELBO} = \mathcal{L}_v^m + \mathcal{L}_t^m. \quad (4)$$

So far, the input heterogeneous data can be encoded into two latent spaces with the modified VAE architecture, then the original features can be reconstructed according to the decoder of the corresponding modality with the principle of loss minimization. Therefore, we can formulate the total multi-latent variable reconstruction loss function for cross-modal retrieval, presented as Eq. 4.

D. Semantic-Shared Representation Learning

The essence of cross-modal retrieval is to capture the semantic consistency of paired heterogeneous data in the common space. Thus, a semantic encoder E_s shared by all modalities is designed to achieve this goal, displayed as Fig. 2. Exceptionally, the first layer of E_s relies on the dimension of the input for the specific modality and transforms them into intermediate representations with the same dimension. Given an image-text pair $o_i = \{v_i, t_i\}$, the features $\{\mathbf{h}_v^i, \mathbf{h}_t^i\}$ with the same dimension can be obtained firstly:

$$\mathbf{h}_v^i = f_{c_v}(v_i); \mathbf{h}_t^i = f_{c_t}(t_i). \quad (5)$$

Then the semantic consistency $\mathbf{z}_v^s \in \mathbb{R}^m$ and $\mathbf{z}_t^s \in \mathbb{R}^m$ can be extracted through the remaining fc layers of shared weights:

$$\mathbf{z}_v^s = E_s(\mathbf{h}_v); \mathbf{z}_t^s = E_s(\mathbf{h}_t). \quad (6)$$

For brevity, the subscript i for the i th instance is omitted in semantic-shared representations. Additionally, the semantic-shared encoder is uniformly formulated as E_s , regardless of the difference at the first fc layer.

To make the extracted semantic representations more similar, additional similarity measures are applied to \mathbf{z}_t^s and \mathbf{z}_v^s . Although compatible with various constraint strategies like P3S [23], we do not need those complex similarity measures and instead utilize the simple, efficient, and robust cross-modal correlation similarity (CMCS) [12], [18], [56], which takes the Euclidean distance to measure semantic consistency of heterogeneous data with their disentangled semantic representations:

$$\mathcal{L}_{sim} = \frac{1}{N} \sum_{i=1}^N \|\mathbf{z}_{v_i}^s - \mathbf{z}_{t_i}^s\|^2. \quad (7)$$

From the perspective of common sense, the extracted semantic representation should be intra-class consistent. Hence, we can align their common representations and thus reduce the discrepancy between them by minimizing this CMCS loss.

The main reasons for adopting this semantic constraint are two-fold: i) We disentangle the latent coupled variables z into two independent variables z^s and z^u , the disentangled semantic-shared representation \mathbf{z}^s can thus get rid of the interference from modality-dependent yet non-semantic representation \mathbf{z}^u . ii) We employ a dual adversarial mechanism, somewhat similar to cross-modal adversary scheme [23], to promote the disentanglement process. The discussion can be referred to in the following subsection III-F.

E. Semantic-Unrelated Representation Learning

In this subsection, we mainly configure an independent encoder for each modality to extract the modality-dependent yet non-semantic representation, on the basis of our modified VAE architecture. Specifically, the semantic-unrelated representations $\mathbf{z}_v^u \in \mathbb{R}^m$ and $\mathbf{z}_t^u \in \mathbb{R}^m$ of the pair-wise instance can be encoded by their own corresponding encoder, displayed as:

$$\mathbf{z}_v^u = E_v(v_i); \mathbf{z}_t^u = E_t(t_i). \quad (8)$$

The re-parameterization trick is also utilized on the semantic-unrelated information \mathbf{z}_*^u . For brevity, we set semantic-shared and -unrelated representations to equal dimensions.

In general, decoders can reconstruct the input representation by combining the shared and specific representations of each modality through concatenation [23], [48]. Nevertheless, our reconstruction by the corresponding decoder is composed of its own modality-dependent representation and the exchanged semantic representation between image and text. It is an intuitive idea that the semantic embeddings of the instances within the same category should preserve consistency regardless of the modalities. Thus, the cross-reconstruction is denoted as:

$$\tilde{v}_i = D_v(\mathbf{z}_v^u \oplus \mathbf{z}_t^s); \tilde{t}_i = D_t(\mathbf{z}_t^u \oplus \mathbf{z}_v^s), \quad (9)$$

where \oplus defines the concatenation operation for the input vectors, \tilde{v}_i and \tilde{t}_i denote the result of reconstruction for v_i and t_i respectively.

Existing methods [12], [17] reconstruct the input only with the representations of a single modality. Their unsatisfactory

performance proves that these common representations are actually entangled with modality-dependent and non-semantic features. Therefore, the disentangled semantic representations can facilitate the correlation capture of heterogeneous data, while the combination of shared and unrelated semantic representations can achieve a robust reconstruction and refactoring. Then, the cross reconstruction objective function of our SDCMR method can be expressed as:

$$\mathcal{L}_{rec} = \frac{1}{N} \sum_{i=1}^N [MSE(\tilde{\mathbf{v}}_i, \mathbf{v}_i) + MSE(\tilde{\mathbf{t}}_i, \mathbf{t}_i)], \quad (10)$$

where $MSE(\cdot)$ denotes Mean Square Error.

F. Representation Disentangling

Our representation disentanglement can be completed through two strategies: dual adversarial mechanism that facilitates the decoupling of semantic-shared and -unrelated representations, and independence constraint that preserves the orthogonality of two disentangled representations.

1) *Dual Adversarial Mechanism*: Displayed as Fig. 2, the structure of **DAM** mainly contains two adversarial units for each modality, and each adversarial unit is shared by all modalities. We refer to the configuration of DANN [57], which is a typical method for domain adaption. As aforementioned, the disentanglement of each modality includes two processes: semantic-shared adversarial learning and semantic-unrelated adversarial learning. The purpose of the former process is to extract the semantic information for the heterogeneous data as much as possible and meanwhile attempt to eliminate all the non-semantic knowledge, which is dependent on the specific modality. Conversely, the latter process struggles to absorb the modality-dependent yet non-semantic knowledge, while makes efforts to exclude the semantic information shared by all the modalities from \mathbf{z}^u .

The function of our semantic-shared adversarial learning is mainly accomplished by a semantic classifier C^s and a modality classifier C^u . To exclude the modality-dependent yet non-semantic information, we configure a gradient reversal layer (GRL) [57] for C^u . Hence, the parameters of semantic-shared encoder E_s are also optimized by minimizing the cross-entropy loss for the semantic classifier C^s while maximizing another cross-entropy loss for the modality classifier C^u . Obviously, the process is a min-max game, like ACMR [16]. Simultaneously, the semantic classifier C^s is also learned during the minimizing optimization. Therefore, the classifiers can promote the discrimination of the semantic-shared and -unrelated representations by semantic-shared adversarial learning, formalized as:

$$\mathcal{L}_{ss} = \frac{1}{N} \sum_{i=1}^N (\text{CE}(C^s(\mathbf{z}_i^s), y_i^s) - \text{CE}(C^u(\mathbf{z}_i^u), y_i^u)) + \frac{1}{N} \sum_{i=1}^N (\text{CE}(C^s(\mathbf{z}_i^s), y_i^s) - \text{CE}(C^u(\mathbf{z}_i^u), y_i^u)), \quad (11)$$

where the $\text{CE}(\cdot)$ operation defines a classical cross-entropy loss, and y_i^s, y_i^u are the corresponding ground truth of the semantic label and modality label for i th instance.

A similar setup holds for the semantic-unrelated adversarial learning process. Differently, the GRL layer is inserted in front of the semantic classifiers C^s to fuse the modality-dependent yet semantic-unrelated knowledge completely. In consequence, the parameters in non-semantic encoder E_* are also effectively optimized by a min-max game, presented as Eq. 11. However, the cross-entropy loss minimization and maximization are exchanged between the semantic classifier C^s and the modality classifier C^u , described as:

$$\mathcal{L}_{su} = \frac{1}{N} \sum_{i=1}^N (\text{CE}(C^u(\mathbf{z}_i^u), y_i^u) - \text{CE}(C^s(\mathbf{z}_i^s), y_i^s)) + \frac{1}{N} \sum_{i=1}^N (\text{CE}(C^u(\mathbf{z}_i^u), y_i^u) - \text{CE}(C^s(\mathbf{z}_i^s), y_i^s)). \quad (12)$$

Thus, the total objective function for our novel dual adversarial mechanism is defined as:

$$\mathcal{L}_{ad} = \mathcal{L}_{su} + \mathcal{L}_{ss}. \quad (13)$$

2) *Independence Constraint*: Moreover, to force the semantic-shared and -unrelated encoders for each modality to learn totally different features from the input, we explicitly add an independence constraint to facilitate the disentanglement through orthogonality operation, borrowed from the P3S method [23]. Specifically, we define two matrices $\{\mathbf{Z}_v^u\}^N \in \mathbb{R}^{N \times m}$ and $\{\mathbf{Z}_t^u\}^N \in \mathbb{R}^{N \times m}$ whose rows are the semantic-unrelated representations of the instances for image and text respectively. Analogously, $\{\mathbf{Z}_v^s\}^N \in \mathbb{R}^{N \times m}$ and $\{\mathbf{Z}_t^s\}^N \in \mathbb{R}^{N \times m}$ are the semantic-shared representations for image and text. Then, the minimization of \mathcal{L}_{ind} loss, shown as Eq. 14, can promote the orthogonality between semantic-shared and unrelated representations of heterogeneous data.

$$\mathcal{L}_{ind} = \left\| (\mathbf{Z}_v^s)^\top \mathbf{Z}_v^u \right\|_F + \left\| (\mathbf{Z}_t^s)^\top \mathbf{Z}_t^u \right\|_F, \quad (14)$$

where $\|\cdot\|_F$ defines the Frobenius norm.

G. Overall Objective and Optimization

To clearly illustrate the optimization of our SDCMR approach, we simplify definitions of the sub-network parameters without ambiguity. Specifically, we employ $\theta_s, \theta_u, \theta_d, \theta_c^s$ and θ_c^u as the simplified parameters of the semantic-shared encoder E_s , the two exclusive encoders E_* for non-semantic knowledge extraction, the two specialized decoders D_* for reconstruction, the semantic classifier C^s and the modality classifier C^u respectively. Then, the parameters are deployed to the corresponding loss, and the final objective function of our SDCMR approach can be derived with a weighted combination of those aforementioned losses.

$$\mathcal{L} = \mathcal{L}_{ELBO} + \alpha \mathcal{L}_{rec} + \beta \mathcal{L}_{sim} + \gamma \mathcal{L}_{ad} + \eta \mathcal{L}_{ind}, \quad (15)$$

where α, β, γ and η are the weights to balance the contribution of each loss.

During the practice, we employ the well-known stochastic gradient descent algorithm, like Adam [58] to optimize the various parameters of Eq. 15 alternately, with μ as the

TABLE I
THE STATISTIC INFORMATION OF THE FOUR DATASETS

Datasets	Classes	Train	Test
Wikipedia	10	2,173	693
Pascal Sentences	20	800	200
NUS-WIDE-10k	10	42941	28661
PKU-XMediaNet	200	32000	8000

learning rate. Firstly, we calculate the gradient of each parameter according to Adam as follows:

$$\theta_u \leftarrow \theta_u - \mu (\alpha \nabla_{\theta_u} \mathcal{L}_{rec} + \eta \nabla_{\theta_u} \mathcal{L}_{ind}), \quad (16)$$

$$\theta_d \leftarrow \theta_d - \mu (\alpha \nabla_{\theta_d} \mathcal{L}_{rec}), \quad (17)$$

$$\theta_c^u \leftarrow \theta_c^u - \mu (\gamma \nabla_{\theta_c^u} \mathcal{L}_{ad}), \quad (18)$$

$$\theta_c^s \leftarrow \theta_c^s - \mu (\gamma \nabla_{\theta_c^s} \mathcal{L}_{ad}), \quad (19)$$

$$\theta_s \leftarrow \theta_s - \mu (\nabla_{\theta_s} \mathcal{L}_{ELBO} + \alpha \nabla_{\theta_s} \mathcal{L}_{rec} + \beta \nabla_{\theta_s} \mathcal{L}_{sim} + \gamma \nabla_{\theta_s} \mathcal{L}_{ad} + \eta \nabla_{\theta_s} \mathcal{L}_{ind}). \quad (20)$$

The detailed training routine of our SDCMR approach is presented in Algorithm 1. During the test, we utilize the learned semantic-shared encoder E_s to extract the semantic-common representations of the tested instances from \mathcal{O}_{te} , i.e., $\{\mathbf{z}_{v_j}^s\}_{j=1}^{N'}$ for the images $\{v_j\}_{j=1}^{N'}$, $\{\mathbf{z}_{t_j}^s\}_{j=1}^{N'}$ for the texts $\{t_j\}_{j=1}^{N'}$ respectively. Then, given an instance of one modality, we can retrieve the semantic manifestation of the corresponding modality by computing the cosine similarities of the two disentangled semantic representations.

IV. EXPERIMENTS

In this section, we conduct comprehensive experiments on four widely used datasets in cross-modal retrieval and make comparisons with 15 state-of-the-art methods to prove the effectiveness and superiority of our proposed SDCMR approach. The performance of our SDCMR method is comprehensively expounded in the following aspects, including experimental settings, results comparison, and further analysis.

A. Experimental Settings

1) *Datasets and Features*: Our cross-modal retrieval experiments are evaluated on three extensively used datasets, i.e. Wikipedia [9], Pascal Sentence [59], NUS-WIDE [60], and a large-scale dataset PKU-XMediaNet [18]. The statistical information of the used datasets is briefly presented in Table I. Note that NUS-WIDE-10k [61] is a cropped subset from NUS-WIDE [60], the original dataset contains 81 classes of approximately 270,000 pairs of image–text data. We pick out a total of 71,602 pairs from the largest 10 classes to make up the subset used in this work. Another thing worth noting is that we adopt the same strategy [17], [62] to select images and text from the PKU-XMediaNet dataset [18], to make our cross-modal retrieval experiments. Namely, the other three modalities, such as audio, video, and 3D model, are not considered in this work.

Following [17], [23], and [62], we adopt VGG-19 [50] as our backbone, and then use the fc -7 layer to extract 4096D

Algorithm 1 The Training Routing of Our SDCMR Approach

Input: Training dataset \mathcal{O}_{tr} , batch size B , learning rate μ , maximum iterations T , hyper-parameters $\alpha, \beta, \gamma, \eta$.

Output: Optimal model parameters: $\theta_s, \theta_u, \theta_d, \theta_c^u, \theta_c^s$.

- 1: **repeat**
- 2: Sample bimodal pairs $\{\mathbf{v}_b, \mathbf{t}_b, y_b\}_{b=1}^B$ from \mathcal{O}_{tr} within a batch.
- 3: Update θ_u based on Eq.16,
- 4: Update θ_d based on Eq.17,
- 5: Update θ_c^u based on Eq.18,
- 6: Update θ_c^s based on Eq.19,
- 7: Update θ_s based on Eq.20.
- 8: **until** The Eq.15 converges or reach maximum iterations.
- 9: The well-learned semantic-shared encoder E_s can disentangle the semantic representations for instances in \mathcal{O}_{te} to perform cross-modal retrieval.

features for each image. For text, the WordCNN [51] is utilized to extract 300D feature of each instance.

2) *Details of Network*: We make our implementation of the SDCMR method with the Pytorch toolkit and the training of the total network is performed with one GeForce RTX 3090 GPU. All encoders (one semantic-shared encoder E_s and two modality-specific encoders E_*) in our SDCMR network are composed of three fc layers. Their dimensions are $[d^* \rightarrow 2048 \rightarrow 1024 \rightarrow k, * = v, t]$, respectively, where k is the dimension for the two disentangled representations. In addition, the *ReLU* activation function follows each fully connected layer. The two exclusive decoders also contain three fc layers yet with a reverse order dimension $[2k \rightarrow 1024 \rightarrow 2048 \rightarrow d^*]$. For the two classifiers, we construct two fc layers of $[k \rightarrow k/2 \rightarrow l_*]$, $\star = u, s$, where l_u and l_s are the categories of modality and semantic label, respectively. The learning rate is set to 0.0002 for Adam optimizer, and we set the mini-batch size and maximum iterations as 64 and 500, respectively. The optimal hyper-parameters are set to 32, 1, 5, 10 and 1 for k, α, β, γ and η respectively.

3) *Comparison Methods*: Our SDCMR method is compared with four types of cross-modal retrieval methods (15 state-of-the-arts in total): 1) five traditional methods including CCA [9], CFA [10], KCCA [27], JRL [11] and LGCFL [63]; 2) five deep-learning (non-GAN) methods namely Corr-AE [12], CMDN [33], Deep-SM [14], MCSM [62] and DSCMR [13]; 3) four GAN-based methods containing ACMR [16], CM-GAN [17], CAAL [42] and TANSS [20]; 4) one representation separation learning methods, that is P3S [23]. For fair comparisons, all methods adopt the same features of image and text as ours. Then, we rank the cosine similarities of the two disentangled semantic representations $\{\mathbf{z}_{v_j}^s\}_{j=1}^{N'}$ and $\{\mathbf{z}_{t_j}^s\}_{j=1}^{N'}$ to complete cross-modal retrieval. Following [17], [23], and [42], we evaluate our performance mainly on two scenarios: 1) Img2Txt is taking a query of image to retrieve the texts describing similar semantic content from \mathcal{O}_{te} ; while 2) the Txt2Img scene aims to obtain images, which can visualize the description of a query in text. Mean average precision (MAP) score, which is extensively used in information retrieval [2], [53], [64]

TABLE II
MAP SCORES OF CROSS-MODAL RETRIEVAL ON FOUR DATASETS FOR THE SDCMR METHOD AND OTHER SOTAS.
THE **BOLD** NUMBER INDICATES THE BEST PERFORMANCE, ALL IMPLEMENTED BY THE PROPOSED METHOD

Methods	Wikipedia			Pascal Sentences			NUS-WIDE-10k			PKU-XMediaNet		
	Img2Txt	Txt2Img	Avg.	Img2Txt	Txt2Img	Avg.	Img2Txt	Txt2Img	Avg.	Img2Txt	Txt2Img	Avg.
CCA [9] (2010)	0.298	0.273	0.286	0.203	0.208	0.206	0.167	0.181	0.174	0.212	0.217	0.215
CFA [10] (2003)	0.319	0.316	0.318	0.476	0.470	0.473	0.406	0.435	0.421	0.252	0.400	0.326
KCCA [27] (2014)	0.438	0.389	0.414	0.488	0.446	0.467	0.351	0.356	0.354	0.252	0.270	0.261
JRL [11] (2014)	0.479	0.428	0.454	0.563	0.505	0.534	0.466	0.499	0.483	0.488	0.405	0.447
LGCFL [63] (2015)	0.466	0.431	0.449	0.539	0.503	0.521	0.453	0.485	0.469	0.441	0.509	0.475
Corr-AE [12] (2014)	0.442	0.429	0.436	0.532	0.521	0.527	0.441	0.494	0.468	0.469	0.507	0.488
CMDN [33] (2016)	0.487	0.427	0.457	0.544	0.526	0.535	0.492	0.542	0.517	0.485	0.516	0.501
Deep-SM [14] (2017)	0.478	0.422	0.450	0.560	0.539	0.550	0.497	0.478	0.488	0.399	0.342	0.371
MCSM [62] (2018)	0.516	0.458	0.487	0.598	0.598	0.598	0.533	0.561	0.547	0.540	0.550	0.545
DSCMR [13] (2019)	0.515	0.479	0.497	0.595	0.598	0.597	0.554	0.563	0.559	0.622	0.632	0.627
ACMR [16] (2017)	0.468	0.412	0.440	0.538	0.544	0.541	0.519	0.542	0.531	0.536	0.519	0.528
CM-GAN [17] (2019)	0.521	0.466	0.494	0.603	0.604	0.604	0.544	0.562	0.553	0.567	0.551	0.559
TANSS [20] (2019)	0.518	0.459	0.488	0.608	0.594	0.601	0.539	0.551	0.545	0.582	0.574	0.578
CAAL [42] (2022)	0.507	0.463	0.485	0.568	0.592	0.580	0.542	0.553	0.548	0.587	0.575	0.581
P3S [23] (2022)	0.520	0.469	0.495	0.622	0.601	0.612	0.545	0.574	0.560	0.647	0.648	0.648
SDCMR (Ours)	0.683	0.475	0.579	0.628	0.608	0.618	0.597	0.607	0.602	0.653	0.661	0.657

and existing cross-modal retrieval methods [17], [23], [42], is also adopted as our evaluation measurement on the two scenarios, as it takes into account both precision and the ranking of all returned results. MAP is calculated as the mean value of average precision for all the queries [23], [48], thus the larger score of MAP manifests the better performance. It is worth noting that MAP scores presented in ACMR [16], Corr-AE [12], DSCMR [13] and CAAL [42] are higher than those in this paper. The reasons could be twofold: 1) the published results are calculated by the top 50 returns for the first three referred methods; 2) The queries for CAAL [42] are not only from the test set, but also from the training set. All MAP scores displayed in this paper are computed with **all** returns in the test set \mathcal{O}_{te} , and we re-evaluate the aforementioned methods with their released codes for fair comparisons.

B. Results Analysis

In Table II, we further showcase the comparison on the two scenarios between our SDCMR approach and other 15 state-of-the-art methods, including five traditional approaches (the 1st panel), five non-GAN based models (the 2nd panel), four GAN based methods (the 3rd panel) and one representation disentangling approach (the last panel).

Dig deep into these MAP scores and we can make some important conclusions:

(1) Deep learning methods, whether non-GAN, GAN-based, or representation disentangling, overall precede traditional methods, due to their strong ability of nonlinear correlation. Only a few methods such as Corr-AE [12] and deep-SM [14] are inferior to traditional approaches like LGCFL [63] and

JRL [11] on PKU-XMediaNet dataset, on account of advanced constraints including local group-based priors and group structure adopted by the latter two methods. With the help of DNNs, our proposed SDCMR approach consistently overcomes all the traditional cross-modal retrieval methods across all datasets, whatever the single scenario or average performance. These results indicate that the SDCMR method can effectively bridge the semantic gap between heterogeneous data.

(2) Thanks to the extremely impressive generation ability, most of the GAN-based methods can exceed non-GAN approaches. However, the advantages of GAN-based methods such as ACMR [16], CM-GAN [17] and TANSS [20] on the large-scale dataset PKU-XMediaNet gradually decrease or even lose. Particularly, they all lag far behind DSCMR, implemented by sharing weights across modalities. This phenomenon is mainly attributed to the complex structure of GAN and its poor generalization ability. To avoid these difficulties, the P3S method [23] based on representation separation, explicitly splits the shared and private subspaces of different modalities to learn a more compact shared subspace. Hence, this separation operation can effectively facilitate the generation of common representations across two modalities, and then remarkably boosts the performance.

(3) When compared to P3S, our SDCMR method takes comprehensive advantages over it. Concretely, in a single retrieval scene, we can lead P3S by a large margin, occurring in the Img2Txt scenario of the Wikipedia dataset, and at least 0.6%, also appearing in Img2Tx on PKU-XMediaNet and Pascal dataset. In addition, on average performance, we outperform the P3S method by at least 0.6% and maximum **8.4%** in each dataset. The possible reason for our improvement is that although the P3S method focuses on separating the

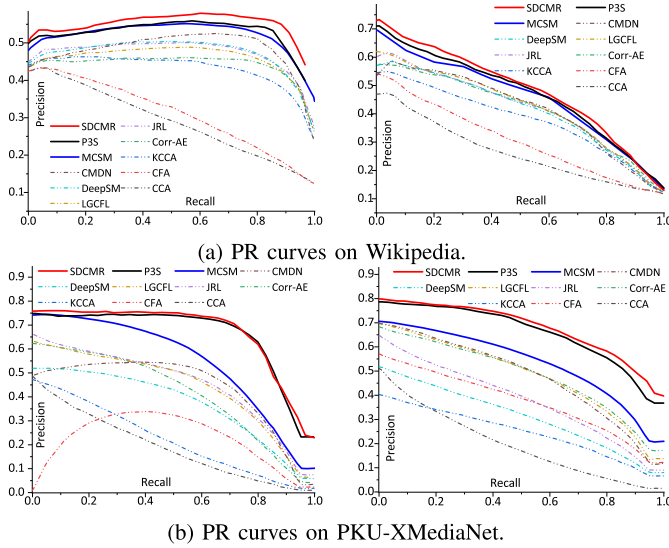









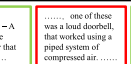



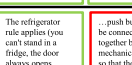

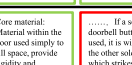



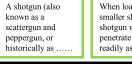
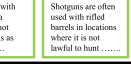



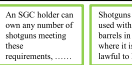
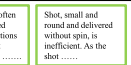
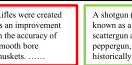


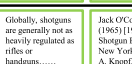
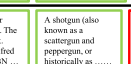



Fig. 3. The precision-recall (PR) curves comparison. The left is for Img2Txt and the right is for Txt2Img, respectively. The top three methods are drawn by solid lines, red for SDCMR, black for P3S, and blue for MCSM, while other methods are dash-dotted lines with different colors.

interference, it may only attend to easy and trivial information, such as irrelevant objects, and fails to completely disentangle the semantic-shared representation. This insufficient attention makes the learned representations partially entangled, which hinders further improvement. The slight improvements on Pascal Sentences and PKU-XMediaNet may be due to their relatively small dataset scale. Additionally, PKU-XMediaNet [18] is better suited for fine-grained tasks while our SDCMR approach does not focus on the variances within classes, which can also affect some performance.



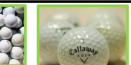
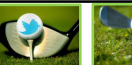


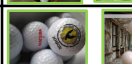

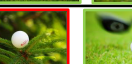


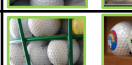
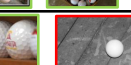
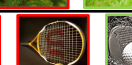


















(4) In summary, the improvement fulfilled by our SDCMR method ascribes great importance to these key factors:

- The VAE structure is utilized to explicitly disentangle the coupled representation of each modality into semantic-shared and -unrelated representations. In this way, the features of semantic correlation for the image-text pair can be better extracted.
- The proposed dual adversarial mechanism effectively promotes the disentanglement through expelling modality-dependent yet non-semantic knowledge from the semantic-shared representation, while squeezing the semantic-related information into the semantic-shared representation.
- In addition, the semantic encoder shared by different modalities can also facilitate the capture of semantic consistent features.

Moreover, Fig. 3 reports the PR curves of the two retrieval scenarios, generated by SDCMR and several comparison methods on Wikipedia and PKU-XMediaNet datasets. It is observed that our SDCMR method coherently keeps the highest accuracy in PR curves regardless of the recall level, which demonstrates the effectiveness of our semantics disentangling strategy. Furthermore, several cross-modal retrieval examples on PKU-XMediaNet obtained by our SDCMR method and other two latest approaches including ACMR and P3S, are visualized in Fig. 4, which demonstrates our superiority again.

Query	Methods	Top five results (Img2Txt)				
 door	SDCMR (Ours)					
	P3S					
	ACMR					
 shotgun	SDCMR (Ours)					
	P3S					
	ACMR					

(a) Examples for Img2Txt.

Query	Methods	Top five results (Txt2Img)				
 Golf ball	SDCMR (Ours)					
	P3S					
	ACMR					
 Seagull	SDCMR (Ours)					
	P3S					
	ACMR					

(b) Examples for Txt2Img.

Fig. 4. Several Img2Txt and Txt2Img retrieval examples on PKU-XMediaNet were obtained by our SDCMR method, and two compared methods P3S [23] and ACMR [16]. The ground-truth of each query is also posted here for further comparison. In addition, the right and incorrect retrieval results are marked with green and red color respectively.

C. Further Analysis

1) *Ablation Study*: Shown as Eq. 15, our final objective function includes five loss items, which promote semantic disentangling together. To investigate the effect of each item, we evaluate the performance of the incomplete method by excluding one loss item every time. As the architecture of our SDCMR method is based on VAE, we do not eliminate the loss function of \mathcal{L}_{ELBO} . For brevity, we only take Wikipedia and NUS-WIDE-10k as our testbeds to perform the ablation study. The results are displayed in Table III, which tell us that each loss item is important to our SDCMR method and makes its own contribution to semantics disentangling.

2) *Comparison With Other Decoupling Methods*: To further demonstrate our superiority, we also transfer the existing disentangling methods including β VAE [22] and SDGZSL [46] to cross-modal retrieval. We still conduct semantic disentangling experiments based on Wikipedia and NUS-WIDE-10k. The comparisons are displayed in Table IV, which re-confirm the superiority and effectiveness of our SDCMR method in semantics disentangling. Besides, multi-variable VAE architecture ("w/o \mathcal{L}_{ad} " in Table III) brings about the performance decline can further demonstrate our clear advantages.

TABLE III
MAP SCORES OF CROSS-MODAL RETRIEVAL ON WIKIPEDIA AND NUS-WIDE-10K DATASETS FOR THE ABLATION STUDY

Methods	Wikipedia			NUS-WIDE-10k		
	Img2Txt	Txt2Img	Avg.	Img2Txt	Txt2Img	Avg.
SDCMR	0.683	0.475	0.579	0.597	0.607	0.602
w/o \mathcal{L}_{rec}	0.675	0.468	0.572	0.578	0.596	0.587
w/o \mathcal{L}_{sim}	0.646	0.410	0.528	0.532	0.562	0.547
w/o \mathcal{L}_{ad}	0.551	0.363	0.457	0.492	0.496	0.494
w/o \mathcal{L}_{ind}	0.599	0.417	0.508	0.568	0.592	0.580

TABLE IV
COMPARISON WITH OTHER DECOUPLING METHODS ON WIKIPEDIA AND NUS-WIDE-10K

Methods	Wikipedia			NUS-WIDE-10k		
	Img2Txt	Txt2Img	Avg.	Img2Txt	Txt2Img	Avg.
SDCMR	0.683	0.475	0.579	0.597	0.607	0.602
SDGZSL	0.542	0.470	0.506	0.563	0.582	0.573
β VAE($\beta = 5$)	0.515	0.471	0.493	0.530	0.563	0.547

TABLE V
INTEGRATION WITH OTHER ADVANCED BACKBONES SUCH AS BLIP-2 ViT-L [65] ON WIKIPEDIA AND NUS-WIDE-10K

Methods	Wikipedia			NUS-WIDE-10k		
	Img2Txt	Txt2Img	Avg.	Img2Txt	Txt2Img	Avg.
BLIP-2	0.804	0.818	0.811	0.820	0.826	0.823
+SDCMR	0.809	0.823	0.816	0.827	0.831	0.829

3) *Integration With Advanced Backbones*: We also employ an advanced backbone as our feature encoder. Specifically, the architecture of BLIP-2 ViT-L [65] based on CLIP [66] is utilized to extract initial feature extraction. Then we integrate our SDCMR method into BLIP V2 for further improvement, displayed as Table V. The results demonstrate our effectiveness, albeit slightly. This is mainly due to the superior performance of BLIP-2.

4) *Parameter Sensitivity*: We also conduct extensive experiments to investigate the sensitivity of hyper-parameters including α , β , γ and η based on Wikipedia datasets. In each experiment, we let one of the parameters vary in the range of [0.001, 1000], while the others remain constant, to roughly search for the optimal value for the best performance. Take the average MAP as an example, the curves of performance as parameters vary are shown in Fig. 5(a), which tells us that the optimal values for α , β , γ and η are 1, 5, 10 and 1, respectively.

Furthermore, the MAP scores (Img2Txt, Txt2Img and average) with error bars for 5 runs of the proposed SDCMR approach on four datasets are displayed in Fig. 5(b), which tells us that the performance of our SDCMR method is robust to the initialization of network.

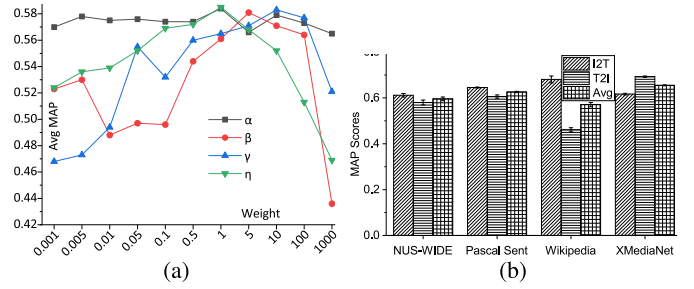


Fig. 5. (a) Parameter sensitivity on Wikipedia. (b) Error bars of MAP scores on different datasets.

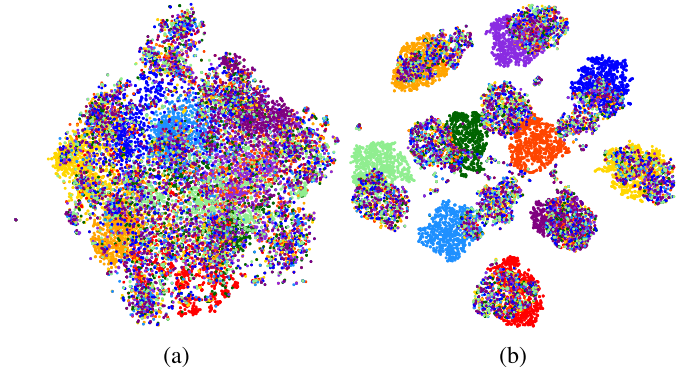


Fig. 6. The t-SNE visualizations of the initial and the disentangled representations on the NUS-WIDE-10k dataset. (a) Initial representations provided by VGG-19 and Word-CNN. (b) Disentangled representations by our SDCMR method.

5) *Visualizations of Semantic Representations Disentangled*: Following the mainstream methods [17], [19], [23], [42] of this field, we also visualize the disentangled semantic representations by our SDCMR method with the t-SNE tool [67] on the NUSWIDE-10k dataset. As a contrast, we also show the initial representation provided by the off-the-shelf backbones, such as VGG-19 [50] and WordCNN [51].

The detailed visualizations are shown in Fig. 6, where the image and text are marked with circles and crosses, respectively. We can make a conclusion that our SDCMR method can disentangle the semantics from the initial entangled representations, and exclude the interference of other nonsemantic information. Therefore, the performance can be beneficial from the discriminative and disentangled semantic representations.

V. CONCLUSION

In this paper, we proposed a semantics disentangling framework for traditional cross-modal retrieval, which utilized the disentanglement of explainable factors based on our modified VAE structure and a novel dual adversarial mechanism to promote the decoupling of semantic-shared and -unrelated representations. Benefiting from the well-disentangled semantic representations, our proposed SDCMR method can circumvent the interference of semantic-unrelated information, and thus improve the performance of cross-modal retrieval. Comprehensive evaluation results on four widely used datasets demonstrate the effectiveness and the superiority of our novel semantics disentangling framework for cross-modal retrieval.

For future work, we will explore more advanced schemes to learn the cross-modal correlation based on our disentangled semantic representations, and extend our semantics disentangling to other modalities besides images and texts.

REFERENCES

- [1] Z. Wang, X. Xu, Y. Zhang, Y. Yang, and H. T. Shen, "Complex relation embedding for scene graph generation," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Dec. 15, 2022, doi: 10.1109/TNNLS.2022.3226871.
- [2] 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.
- [3] E. Yu, J. Sun, J. Li, X. Chang, X. Han, and A. G. Hauptmann, "Adaptive semi-supervised feature selection for cross-modal retrieval," *IEEE Trans. Multimedia*, vol. 21, no. 5, pp. 1276–1288, May 2019.
- [4] Z. Wang, Z. Gao, X. Xu, Y. Luo, Y. Yang, and H. T. Shen, "Point to rectangle matching for image text retrieval," in *Proc. 30th ACM Int. Conf. Multimedia*, Oct. 2022, pp. 4977–4986.
- [5] Z. Wang, X. Xu, J. Wei, N. Xie, J. Shao, and Y. Yang, "Quaternion representation learning for cross-modal matching," *Knowl.-Based Syst.*, vol. 270, Jun. 2023, Art. no. 110505.
- [6] Z. Zhang et al., "Scalable supervised asymmetric hashing with semantic and latent factor embedding," *IEEE Trans. Image Process.*, vol. 28, no. 10, pp. 4803–4818, Oct. 2019.
- [7] E. Yu, J. Ma, J. Sun, X. Chang, H. Zhang, and A. G. Hauptmann, "Deep discrete cross-modal hashing with multiple supervision," *Neurocomputing*, vol. 486, pp. 215–224, May 2022.
- [8] Z. Wang, X. Xu, G. Wang, Y. Yang, and H. T. Shen, "Quaternion relation embedding for scene graph generation," *IEEE Trans. Multimedia*, vol. 25, pp. 8646–8656, 2023.
- [9] N. Rasiwasia et al., "A new approach to cross-modal multimedia retrieval," in *Proc. 18th ACM Int. Conf. Multimedia*, Oct. 2010, pp. 251–260.
- [10] D. Li, N. Dimitrova, M. Li, and I. K. Sethi, "Multimedia content processing through cross-modal association," in *Proc. 11th ACM Int. Conf. Multimedia*, Nov. 2003, pp. 604–611.
- [11] X. Zhai, Y. Peng, and J. Xiao, "Learning cross-media joint representation with sparse and semisupervised regularization," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 6, pp. 965–978, Jun. 2014.
- [12] F. Peng, X. Wang, and R. Li, "Cross-modal retrieval with correspondence autoencoder," in *Proc. 22nd ACM Int. Conf. Multimedia*, Sep. 2014, pp. 7–16.
- [13] 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. 10386–10395.
- [14] Y. Wei et al., "Cross-modal retrieval with CNN visual features: A new baseline," *IEEE Trans. Cybern.*, vol. 47, no. 2, pp. 449–460, Feb. 2017.
- [15] P. Hu, L. Zhen, D. Peng, and P. Liu, "Scalable deep multimodal learning for cross-modal retrieval," in *Proc. 42nd Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2019, pp. 635–644.
- [16] B. Wang, Y. Yang, X. Xu, A. Hanjalic, and H. T. Shen, "Adversarial cross-modal retrieval," in *Proc. 25th ACM Int. Conf. Multimedia*, Oct. 2017, pp. 154–162.
- [17] Y. Peng and J. Qi, "CM-GANs: Cross-modal generative adversarial networks for common representation learning," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 15, no. 1, pp. 1–24, Feb. 2019.
- [18] X. Huang, Y. Peng, and M. Yuan, "MHTN: Modal-adversarial hybrid transfer network for cross-modal retrieval," *IEEE Trans. Cybern.*, vol. 50, no. 3, pp. 1047–1059, Mar. 2020.
- [19] X. Xu, K. Lin, Y. Yang, A. Hanjalic, and H. T. Shen, "Joint feature synthesis and embedding: Adversarial cross-modal retrieval revisited," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 6, pp. 3030–3047, Jun. 2022.
- [20] X. Xu, H. Lu, J. Song, Y. Yang, H. T. Shen, and X. Li, "Ternary adversarial networks with self-supervision for zero-shot cross-modal retrieval," *IEEE Trans. Cybern.*, vol. 50, no. 6, pp. 2400–2413, Jun. 2020.
- [21] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1798–1828, Aug. 2013.
- [22] I. Higgins et al., " β -VAE: Learning basic visual concepts with a constrained variational framework," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2017, pp. 1–10.
- [23] X. Xu, K. Lin, L. Gao, H. Lu, H. T. Shen, and X. Li, "Learning cross-modal common representations by private-shared subspaces separation," *IEEE Trans. Cybern.*, vol. 52, no. 5, pp. 3261–3275, May 2022.
- [24] Y. Zhuang, Y. Wang, F. Wu, Y. Zhang, and W. Lu, "Supervised coupled dictionary learning with group structures for multi-modal retrieval," in *Proc. 27th AAAI Conf. Artif. Intell. (AAAI)*, 2013, pp. 1070–1076.
- [25] Y. Gong, Q. Ke, M. Isard, and S. Lazebnik, "A multi-view embedding space for modeling internet images, tags, and their semantics," *Int. J. Comput. Vis.*, vol. 106, no. 2, pp. 210–233, Jan. 2014.
- [26] K. Wang, R. He, L. Wang, W. Wang, and T. Tan, "Joint feature selection and subspace learning for cross-modal retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 10, pp. 2010–2023, Oct. 2016.
- [27] 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, Dec. 2004.
- [28] V. Ranjan, N. Rasiwasia, and C. V. Jawahar, "Multi-label cross-modal retrieval," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4094–4102.
- [29] M. Xu, Z. Zhu, X. Zhang, Y. Zhao, and X. Li, "Canonical correlation analysis with $L_{2,1}$ -norm for multiview data representation," *IEEE Trans. Cybern.*, vol. 50, no. 11, pp. 4772–4782, Nov. 2020.
- [30] K. Wang, R. He, W. Wang, L. Wang, and T. Tan, "Learning coupled feature spaces for cross-modal matching," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 2088–2095.
- [31] D. Wang, X. 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. 2019.
- [32] Z. Wang, X. Xu, Y. Luo, G. Wang, and Y. Yang, "Hypercomplex context guided interaction modeling for scene graph generation," *Pattern Recognit.*, vol. 141, Sep. 2023, Art. no. 109634.
- [33] Y. Peng, X. Huang, and J. Qi, "Cross-media shared representation by hierarchical learning with multiple deep networks," in *Proc. Int. Joint Conf. Artif. Intell.*, 2016, pp. 3846–3853.
- [34] Y. Peng, J. Qi, X. Huang, and Y. Yuan, "CCL: Cross-modal correlation learning with multigrained fusion by hierarchical network," *IEEE Trans. Multimedia*, vol. 20, no. 2, pp. 405–420, Feb. 2018.
- [35] X. Huang, Y. Peng, and M. Yuan, "Cross-modal common representation learning by hybrid transfer network," in *Proc. 26th Int. Joint Conf. Artif. Intell.*, Aug. 2017, pp. 1893–1900.
- [36] P. Kang, Z. Lin, Z. Yang, X. Fang, Q. Li, and W. Liu, "Deep semantic space with intra-class low-rank constraint for cross-modal retrieval," in *Proc. Int. Conf. Multimedia Retr.*, Jun. 2019, pp. 226–234.
- [37] Y. Wu, S. Wang, G. Song, and Q. Huang, "Online asymmetric metric learning with multi-layer similarity aggregation for cross-modal retrieval," *IEEE Trans. Image Process.*, vol. 28, no. 9, pp. 4299–4312, Sep. 2019.
- [38] Y. Zhang, W. Zhou, M. Wang, Q. Tian, and H. Li, "Deep relation embedding for cross-modal retrieval," *IEEE Trans. Image Process.*, vol. 30, pp. 617–627, 2021.
- [39] I. J. Goodfellow et al., "Generative adversarial nets," in *Proc. 27th Int. Conf. Neural Inf. Process. Syst.* Cambridge, MA, USA: MIT Press, 2014, pp. 2672–2680.
- [40] Y. Lu, S. Wu, Y.-W. Tai, and C.-K. Tang, "Image generation from sketch constraint using contextual GAN," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 213–228.
- [41] S. Pidhorskyi, D. A. Adjeroh, and G. Doretto, "Adversarial latent autoencoders," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 14092–14101.
- [42] S. He et al., "Category alignment adversarial learning for cross-modal retrieval," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 5, pp. 4527–4538, May 2023.
- [43] C. Gong, "Exploring commonality and individuality for multi-modal curriculum learning," in *Proc. 31st AAAI Conf. Artif. Intell. (AAAI)*, 2017, pp. 1926–1933.
- [44] M. Jia, X. Cheng, S. Lu, and J. Zhang, "Learning disentangled representation implicitly via transformer for occluded person re-identification," *IEEE Trans. Multimedia*, vol. 25, pp. 1294–1305, 2023.
- [45] G. Wang, C. Sun, X. Xu, J. Li, Z. Wang, and Z. Ma, "Disentangled representation learning and enhancement network for single image de-raining," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 3015–3023.
- [46] Z. Chen et al., "Semantics disentangling for generalized zero-shot learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 8692–8700.

- [47] J. Y. Huang, K.-H. Huang, and K.-W. Chang, "Disentangling semantics and syntax in sentence embeddings with pre-trained language models," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2021, pp. 1372–1379.
- [48] F. Wu et al., "Modality-specific and shared generative adversarial network for cross-modal retrieval," *Pattern Recognit.*, vol. 104, Aug. 2020, Art. no. 107335.
- [49] J. Deng, W. Ou, J. Gou, H. Song, A. Wang, and X. Xu, "Representation separation adversarial networks for cross-modal retrieval," *Wireless Netw.*, vol. 3, pp. 1–14, Jun. 2020.
- [50] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2015, pp. 1–10.
- [51] Y. Kim, "Convolutional neural networks for sentence classification," in *Proc. Conf. Empir. Method Nat. Lang. Process. (EMNLP)*, 2014, pp. 1746–1751.
- [52] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2014, pp. 1–10.
- [53] K. Lin, X. Xu, L. Gao, Z. Wang, and S. H. Tao, "Learning cross-aligned latent embeddings for zero-shot cross-modal retrieval," in *Proc. 34th AAAI Conf. Artif. Intell. (AAAI)*, 2020, pp. 11515–11522.
- [54] X. Zhu, C. Xu, and D. Tao, "Where and what? Examining interpretable disentangled representations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 5857–5866.
- [55] Y. Zou, X. Yang, Z. Yu, B. V. Kumar, and J. Kautz, "Joint disentangling and adaptation for cross-domain person re-identification," in *Proc. 16th Eur. Conf. Comput. Vis. (ECCV)*. Glasgow, U.K.: Springer, Aug. 2020, pp. 87–104.
- [56] X. Xu, J. Song, H. Lu, Y. Yang, F. Shen, and Z. Huang, "Modal-adversarial semantic learning network for extendable cross-modal retrieval," in *Proc. ACM Int. Conf. Multimedia Retr.*, Jun. 2018, pp. 46–54.
- [57] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by back-propagation," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 1180–1189.
- [58] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Int. Conf. Learn. Represent. (ICLR)*, 2014, pp. 1–10.
- [59] C. Rashtchian, P. Young, M. Hodosh, and J. Hockenmaier, "Collecting image annotations using Amazon's mechanical Turk," *Proc. Workshop Creating Speech Lang. Data With Amazon's Mechanical Turk*, 2010, pp. 139–147.
- [60] 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.*, Jul. 2009, pp. 1–9.
- [61] X. Zhai, Y. Peng, and J. Xiao, "Heterogeneous metric learning with joint graph regularization for cross-media retrieval," in *Proc. 27th AAAI Conf. Artif. Intell. (AAAI)*, 2013, pp. 1198–1204.
- [62] Y. Peng, J. Qi, and Y. Yuan, "Modality-specific cross-modal similarity measurement with recurrent attention network," *IEEE Trans. Image Process.*, vol. 27, no. 11, pp. 5585–5599, Nov. 2018.
- [63] C. Kang, S. Xiang, S. Liao, C. Xu, and C. Pan, "Learning consistent feature representation for cross-modal multimedia retrieval," *IEEE Trans. Multimedia*, vol. 17, no. 3, pp. 370–381, Mar. 2015.
- [64] Z. Zhang, H. Luo, L. Zhu, G. Lu, and H. T. Shen, "Modality-invariant asymmetric networks for cross-modal hashing," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 5, pp. 5091–5104, May 2023.
- [65] J. Li, D. Li, S. Savarese, and S. Hoi, "BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," 2023, *arXiv:2301.12597*.
- [66] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 8748–8763.
- [67] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 86, pp. 2579–2605, 2008.

Zheng Wang received the Ph.D. degree from Zhejiang University, China, in 2017. He is currently with the School of Computer Science and Engineering, University of Electronic Science and Technology of China, China. His current research interests include multimedia understanding and computer vision.

Xing Xu (Member, IEEE) received the Ph.D. degree from Kyushu University, Japan, in 2015. He is currently with the School of Computer Science and Engineering, University of Electronic Science and Technology of China, China. He is a recipient of six academic awards, including the IEEE Multimedia Prize Paper 2020, the Best Paper Award from ACM Multimedia 2017, and the World's FIRST 10K Best Paper Award-Platinum Award from IEEE ICME 2017. His current research interests include multimedia information retrieval and computer vision.

Jiwei Wei received the Ph.D. degree from the University of Electronic Science and Technology of China (UESTC), under the supervision of Prof. Yang Yang, in January 2022. In January 2022, he joined the School of Computer Science and Engineering, UESTC, as a Postdoctoral Research Fellow. His current research interests include multimedia information retrieval, metric learning, and computer vision.

Ning Xie received the M.E. and Ph.D. degrees from the Department of Computer Science, Tokyo Institute of Technology, Tokyo, Japan, in 2009 and 2012, respectively. In 2012, he was appointed as a Research Associate with Tokyo Institute of Technology. Since 2017, he has been an Associate Professor with the Center for Future Media, School of Computer Science and Engineering, University of Electronic Science and Technology of China (UESTC). His research interests include computer graphics, game engine, and the theory and application of artificial intelligence and machine learning. His research is supported by research grants, including NSFC, China, MOE, China, CREST, Japan, and The Ministry of Education, Culture, Sports, Science and Technology, Japan.

Yang Yang (Senior Member, IEEE) received the Ph.D. degree in computer science from The University of Queensland, Brisbane, QLD, Australia, in 2012. He is currently with the University of Electronic Science and Technology of China, Chengdu, China. He was a Research Fellow with the National University of Singapore, Singapore, from 2012 to 2014. His current research interests include multimedia content analysis, computer vision, and social media analytics.

Heng Tao Shen (Fellow, IEEE) received the B.Sc. degree (Hons.) and the Ph.D. degree from the Department of Computer Science, National University of Singapore, in 2000 and 2004, respectively. He is currently the Dean of the School of Computer Science and Engineering and the Executive Dean of the AI Research Institute, University of Electronic Science and Technology of China (UESTC). His research interests mainly include multimedia search, computer vision, artificial intelligence, and big data management. He is/was an Associate Editor of *ACM Transactions of Data Science*, *IEEE TRANSACTIONS ON IMAGE PROCESSING*, *IEEE TRANSACTIONS ON MULTIMEDIA*, *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING*, and *Pattern Recognition*. He is a fellow of ACM and OSA.