Exploring Graph-Structured Semantics for Cross-Modal Retrieval

Lei Zhang, Leiting Chen, Chuan Zhou*†
School of Computer Science and Engineering, University of Electronic Science and Technology of China
Chengdu, China
lei_zhang@std.uestc.edu.cn
{richardchen,zhouchuan}@uestc.edu.cn

Fan Yang, Xin Li
AIQ
Abu Dhabi, UAE
{fanyang_uestc,xinli_uestc}@hotmail.com

ABSTRACT

We study and address the cross-modal retrieval problem which lies at the heart of visual-textual processing. Its major challenge lies in how to effectively learn a shared multi-modal feature space where the discrepancies of semantically related pairs, such as images and texts, are minimized regardless of their modalities. Most current methods focus on reasoning about cross-modality semantic relations within individual image-text pair to learn the common representation. However, they overlook more global, structural inter-pair knowledge within the dataset, *i.e.,* the graph-structured semantics within each training batch. In this paper, we introduce a graph-based, semantic-constrained learning framework to comprehensively explore the intra- and inter-modality information for cross-modal retrieval. Our idea is to maximally explore the structures of labeled data in graph latent space, and use them as semantic constraints to enforce feature embeddings from the semantically-matched (image-text) pairs to be more similar and vice versa. It raises a novel graph-constrained common embedding learning paradigm for cross-modal retrieval, which is largely under-explored up to now. Moreover, a GAN-based dual learning approach is used to further improve the discriminability and model the joint distribution across different modalities. Our fully-equipped approach, called Graph-constrained Cross-modal Retrieval (GCR), is able to mine intrinsic structures of training data for model learning and enable reliable cross-modal retrieval. We empirically demonstrate that our GCR can achieve higher accuracy than existing state-of-the-art approaches on Wikipedia, NUS-WIDE-10K, PKU XMedia and Pascal Sentence datasets. Our code will be made publicly available. Code is available at https://github.com/neoscheung/GCR.

*Chuan Zhou is the corresponding author.
†Lei Zhang, Leiting Chen and Chuan Zhou are also affiliated with Digital Media Technology Key Laboratory of Sichuan Province, University of Electronic Science and Technology of China, Chengdu, China. Leiting Chen and Chuan Zhou are also affiliated with Institute of Electronic and Information Engineering of UESTC in Guangdong, Dongguan, China.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
MM '21, October 20–24, 2021, Chengdu, Sichuan Province, China.
© 2021 Association for Computing Machinery.
ACM ISBN 978-1-4503-8651-7/21/10…$15.00
https://doi.org/10.1145/3474085.3475567

CCS CONCEPTS

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

KEYWORDS

cross-modal retrieval; adversarial learning; graph neural network

ACM Reference Format:
Lei Zhang, Leiting Chen, Chuan Zhou and Fan Yang, Xin Li. 2021. Exploring Graph-Structured Semantics for Cross-Modal Retrieval. In *Proceedings of the 29th ACM Int'l Conference on Multimedia (MM '21), Oct. 20–24, 2021, Chengdu, Sichuan Province, China.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3474085.3475567

1 INTRODUCTION

Cross-modal retrieval (CMR), which aims to search the semantic relevant samples across different modalities, is a fundamental problem in the field of multimedia. It plays a crucial role in a wide variety of high-level AI applications, such as visual captioning [1, 29], visual-question answering [16, 18] and visual grounding [8].

The main difficulty of CMR comes from the 'heterogeneity gap' between the queries and the retrieval candidates of other modalities. Traditional methods tend to tackle this issue by employing the statistical correlation analysis for learning the common subspace, *e.g.,* canonical correlation analysis [6, 10, 14, 28], partial least squares [4, 30] or bilinear model [31, 33]. Most of traditional approaches belong to linear-projection methods, which cannot fully model the correlation of cross-modal data in real-world scenarios. Recent deep learning based methods[9, 15, 22, 23, 25, 40, 48] transform data of multiple sources to the common space with end-to-end learnable projections, yielding a more powerful, highly non-linear solution. Basically, these approaches utilize non-linear neural network to model the mapping functions between inputs and embeddings, and mine the inter-modality semantic relevance to achieve a consistent representation.

Despite the great success of deep learning techniques in CMR, the standard deep learning solutions generally suffer from two limitations. **First**, a lot of approaches focus primarily on the inter-modal relations / relevances, ignoring the intra-modal context of the whole dataset during training, *i.e.,* the correlations across images or texts. **Second**, the structures of labeled data, which reflect the intrinsic relations and prior knowledge, are not fully exploited by existing approaches. That is, current solutions only utilize the labeled data to blindly pull semantic relevant ones together or push those mismatched ones away, while overlook the more valuable structural information. Hence, the trained model may fail to achieve optimal

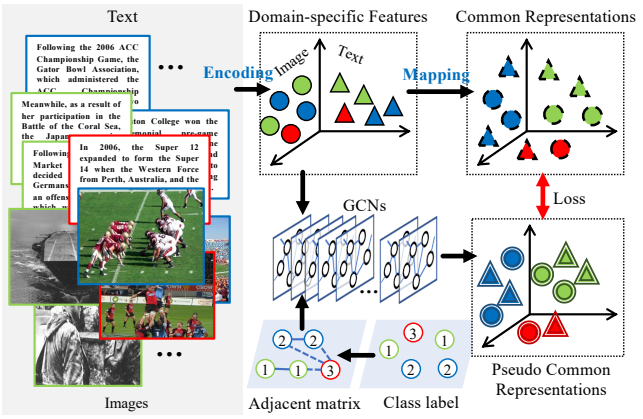

Figure 1: Idea Illustration. Our idea is to explore structural semantics / priors across each individual sample to create the pseudo common representations as extra constraints for common representation learning. Best viewed in color.

latent representations that can fully reflect the intrinsic structural semantics of data. From a higher perspective, the collection of labeled, multi-modality data as a whole contains stronger, structural semantics beyond each single sample that can mitigate the ambiguity in cross-modal retrieval. Yet, how to effectively mine the graph-structured semantics for learning (cross-modality) semantic embeddings is still largely under-explored.

In this paper, we rethink the de facto learning paradigm for cross-modal retrieval by further considering structural semantics within the dataset. As illustrated in Figure 1, our idea is to leverage groundtruth labels to build structural relation graphs for creating pseudo common embeddings, and treat them as extra constraints for common representation learning. To achieve this, we design a novel Graph-constrained Structure Encoding (GSE) module to explicitly incorporate the underlying structural (semantic) priors to build the pseudo common embeddings. Specifically, GSE consists of a two-branch graph convolutional network (GCN) for learning the pseudo embeddings of each modality, and a shared classifier to enforce the pseudo embeddings of each modality to be consistent across modalities. In particular, within our GSE, we build the underlying graphs by linking semantically similar training data within each modality (images or texts) and utilize GCN layers to reason about their intra-modality relevances. In this way, intrinsic structures of training data can be fully explored to build discriminative pseudo embeddings, which encodes semantics beyond each individual image / text. Moreover, we feed the learned pseudo embeddings to a shared classifier, such that we can use classification labels as supervision signals to encourage consistent representations for positive pairs, and inconsistent representations for negative ones, regardless of their modalities. As the pseudo embeddings are created by encoding all structural priors of the dataset derived from manual annotations, they can be considered as another form of supervision signals for learning the intra-modality feature discriminativeness. Thereby, we utilize Frobenius norm to include them as important constraints for training the common feature projector, with the aim of exploring graph-structured semantics and shaping the embedding space.

Along with GSE, we build our feature projectors for common feature extraction as a GAN-based Dual Learning network (GDL). Generally, our idea is to inject dual learning [13] into adversarial representation learning [42]. Specifically, our GDL contains two generative adversarial networks (GANs), in which the generative sub-networks act as common feature projectors to fit the joint distribution and the discriminative sub-networks learn to judge the relevance between different modalities. It allows the common feature projectors model the joint distribution across modalities in an unsupervised manner.

Our fully-equipped approach, called Graph-constrained Cross-modal Retrieval (GCR), integrates GSE and GDL within a unified framework for learning the common representations. To train our GCR, we employ a symmetrical training algorithm, which iterates over two modalities to achieve the desired model parameters of feature projectors. We demonstrate that our GCR can achieve state-of-the-art (SOTA) performance on Wikipedia, NUS-WIDE-10K, PKU XMedia and Pascal Sentence datasets. The contributions of this work are summarized as follows:

- **A new graph-constrained common representation learning approach.** We propose a graph-based, semantic-constrained learning approach to explore the intra-modality and inter-modality information for cross-modal retrieval. It goes beyond each individual image / text and mines the graph-structured semantics among labeled data for building the common feature embeddings.
- **A novel GAN-based dual learning framework.** We design our feature projectors as a GAN-based dual learning network. It interweaves the projectors of different modalities by two discriminative models for learning the joint distribution. It enables each generative model to guide the representation learning of other modality in an unsupervised manner.
- **State-of-the-art results on widely-used benchmarks.** With our symmetrical training algorithm, our fully-equipped GCR sets new records on multiple datasets, *i.e.,* Wikipedia, NUS-WIDE-10K, PKU XMedia and Pascal Sentence datasets, and outperforms existing CMR approaches by a large margin.

2 RELATED WORK

Cross-Modal Retrieval A variety of approaches have been proposed to handle the cross-modal retrieval (CMR) problem. Traditional approaches for CMR mainly based on linear model to generate a lower-dimensional common space with handcrafted or learnable features [4, 6, 10, 12, 14, 28, 30, 31, 33] for multimedia data. Recently, deep neural network (DNN) based approaches learn the common representations with end-to-end learnable projections [9, 15, 22, 23, 25, 40, 48]. As the deep learning approaches have strong ability of non-linear correlation modeling, they can learn the common space more reliably. Ngiam et al. [22] proposed a bi-modal autoencoder approach to model the cross-modal common representation at the shared layer.Feng et al. [9] presented a correspondence autoencoder (Corr-AE) method, which jointly models the cross-modal correlation and reconstruction information to learn projectors. Basically, these methods utilize neural network to extract the feature of single modality and a joint layer to correlate the data across modalities. To better leverage the intra- and inter-modality relevances, Peng et al. [23] proposed cross-modal multiple

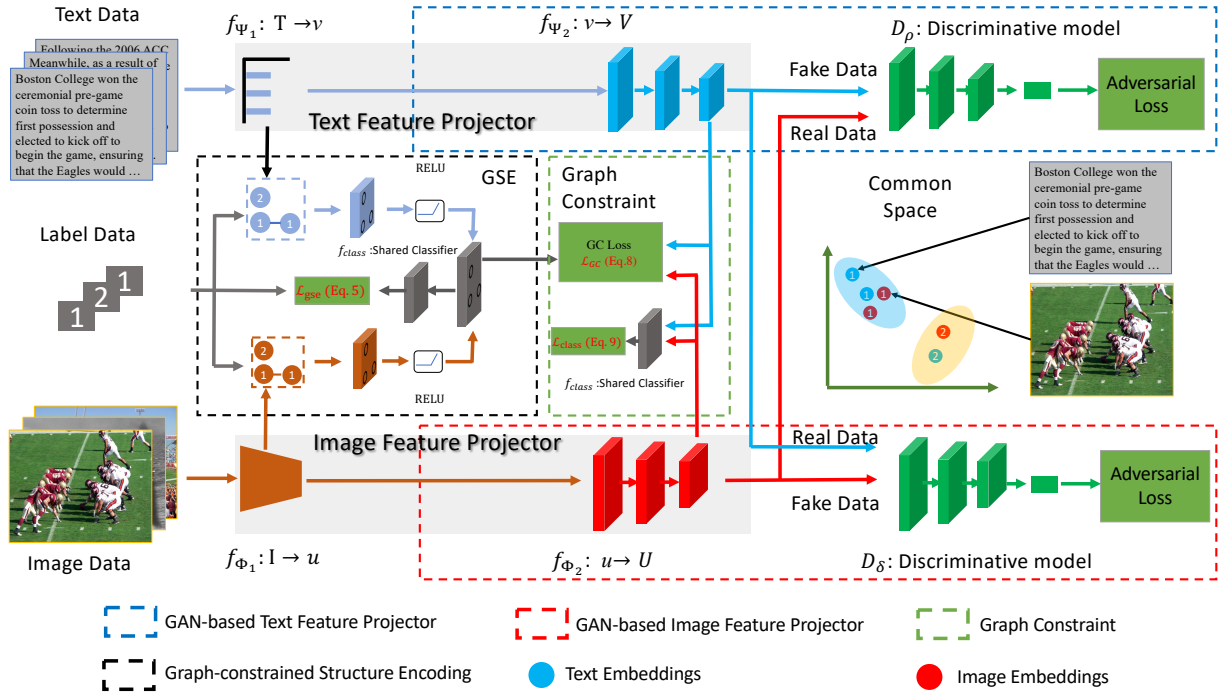

Figure 2: Detailed illustration of GCR. Our GCR includes three major components: the domain specific feature projectors, the Graph-constrained Structure Encoding (GSE) module and the GAN-based Dual Learning (GDL) network. Once trained, only the image & text feature projectors will be used in the inference stage. Please refer to § 3 for more details. Best viewed in color.

deep networks (CMDN) to learn the common representation by hierarchical learning. Zhen et al. [48] proposed deep supervised cross-modal retrieval (DSCMR), which minimizes the discrimination loss in both the label space and the common representation space so as to train the projection networks. To further model the joint distribution over the heterogeneous data, adversarial learning [11] have been widely used in cross-modal retrieval. Wang et al. [35] introduced an adversarial cross-modal retrieval (ACMR) method, for the first time, to learn modality-invariant representations. Peng et al. [24] proposed cross-modal generative adversarial networks (CM-GANs) to model the joint distribution over the data of different modalities. Though these methods can fit the joint distribution well, they are weak in summarizing the high-level semantics within images / texts, because they do not explicitly explore the intra-modality relations. In this paper, we introduce a new deep learning approach for CMR, which can best mine the intra- and inter-modality relevances for building the common representation by combining graph- and GAN-related techniques.

Graph Convolutional Networks. Graph convolutional network (GCN) [3] has been widely used for handling the underlying relationships among structured data [20, 21, 36, 38, 41, 44, 49]. In the context of cross-modal retrieval, GCNs are used to learn common representations across different modalities. Xu et al. [39] proposed graph convolutional hashing model to learn modality-unified binary codes for cross-modal retrieval. Yu et al. [43] utilized GCN to mine relations over each text for learning more reliable text representation. Unlike these existing graph-based approaches, our GCR focuses on reasoning over intra-modality relations to create pseudo common embeddings. In this way, the graph-structured semantics beyond the individual image / text can be fully explored for learning discriminative common embeddings of high-level semantics. To the best of our knowledge, using GCNs to mine structural semantics of training data for providing supervision signals is new for CMR and related domains, which has never been explored before.

3 METHODOLOGY

3.1 Preliminaries

Task Setup and Notations. To begin with, we provide a formal definition of cross-modal retrieval (CMR) task. Here, we focus on the image-text retrieval problem. Let $\mathcal{D} = (\mathcal{I}, \mathcal{T}, \mathcal{Y})$ denote a image-text training set, where $\mathcal{I} = \{\mathbf{I}_i\}_{i=1}^n$ is a set of images, $\mathcal{T} = \{\mathbf{T}_i\}_{i=1}^n$ means a set of the corresponding descriptive sentences, and $\mathcal{Y} = \{\mathbf{Y}_i\}_{i=1}^n$ denotes a collection of associated semantic label vectors. Given the training set $\mathcal{D}$, our task is to learn two mapping functions (feature projectors) for two modalities: $U_i = f_\Phi(\mathbf{I}_i) \in \mathbb{R}^d$ for images and $V_i = f_\Psi(\mathbf{T}_i) \in \mathbb{R}^d$ for texts, where d means the dimensionality of the representation in the common representation space, and Φ and Ψ are the learnable parameters. The ideal mapping functions $f_\Phi(\mathbf{I}_i)$ and $f_\Psi(\mathbf{T}_i)$ should be able to create modality-invariant, semantic discriminative representations that can be easily matched across modalities.

Our Idea. Unlike prior arts [15, 22, 23, 25, 48] that explore cross-modality semantic relations within individual image-text pair to learn the common representation, our idea is to leverage the structural semantics within the data to learn semantic discriminative yet modality-invariant representations. We argue that fully mining graph-structured information of the dataset can significantly

improve the feature discriminativeness and better capture the high-level semantics of each sample (image or text), which would help to shape the common embeddings for better handling the CMR task.

Approach Overview. To verify our idea, we present a novel Graph-constrained Cross-modal Retrieval (GCR) approach. As shown in Figure 2, our GCR consists of three major components:

• **Feature Projectors (§ 3.2).** Similar to existing deep learning approaches, we parameterize the feature projectors, f_Φ and f_Ψ, by neural networks. For f_Φ, we consider it as a two-stage process: f_{Φ_1} first takes the RGB image $\mathbf{I} \in \mathbb{R}^{H\times W\times 3}$ as input, and produces domain-specific features $u \in \mathbb{R}^d$: $\mathbf{I} \xrightarrow{f_{\Phi_1}} u$, and then f_{Φ_2} further maps u to the common feature representation $U \in \mathbb{R}^{H\times W\times d}$: $u \xrightarrow{f_{\Phi_2}} U$. As for f_Ψ, we also utilize such two-stage mapping strategy. In the first stage, f_{Ψ_1} takes the original text $\mathbf{T}$ and generates its corresponding domain-specific features $v \in \mathbb{R}^d$: $\mathbf{T} \xrightarrow{f_{\Psi_1}} v$. In the second stage, f_{Ψ_2} further maps v to common features V: $v \xrightarrow{f_{\Psi_2}} V$.

• **Graph-constrained Structure Encoding Module (§ 3.3).** As the key component of our GCR, the Graph-constrained Structure Encoding (GSE) module f_χ is used to produce the pseudo common feature embeddings $\check{\mathbf{U}}$ or $\check{\mathbf{V}}$ for each modality. We design our GSE f_χ as a two-branch graph convolutional network (GCN) with a shared classifier. To comprehensively encode semantics across each individual sample, it takes multiple samples of data (domain-specific features) within each training batch as inputs (*i.e.,* $\{u_1, \ldots, u_k\}$ and $\{v_1, \ldots, v_k\}$) and uses manual labels as prior knowledge for constructing the underlying graph $\mathcal{G}$ to learn the pseudo common feature embeddings $\check{\mathbf{U}}$ or $\check{\mathbf{V}}$: $\{u_i\}_{i=1}^k \xrightarrow{GCN_u} \{\check{\mathbf{U}}_i\}_{i=1}^k$ or $\{v_i\}_{i=1}^k \xrightarrow{GCN_v} \{\check{\mathbf{V}}_i\}_{i=1}^k$. To encourage modality-invariant yet semantic discriminative representations, a shared classifier is used to further transfer $\check{\mathbf{U}}$ or $\check{\mathbf{V}}$ to the corresponding prediction $\check{\mathbf{C}}_u$ or $\check{\mathbf{C}}_v$ for calculating the loss, supervised by using the classification labels. Note that the generated pseudo embeddings ($\check{\mathbf{U}}$ and $\check{\mathbf{V}}$) are used as important semantic constraints for training the feature projectors.

• **GAN-based Dual Learning Network (§ 3.4).** To better fit the joint distribution across modalities, we further upgrade the second stage of our feature projectors by adding two discriminative sub-networks, which forms a GAN-based Dual Learning (GDL) network. In this way, we can treat the common feature projector of each modality as the generator model for producing feature embeddings: $u \xrightarrow{f_{\Phi_2}} U$ or $v \xrightarrow{f_{\Psi_2}} V$; And the newly-added discriminator model D_δ (or D_ρ) is used to estimate the probability that a sample came from f_{Ψ_2} (or f_{Φ_2}) rather than U (or V).

We integrate above three major components together to form a unified learning framework (Sec. 3.5), called Graph-constrained Cross-modal Retrieval (GCR). Our GCR is trained by using a novel symmetrical training algorithm with multiple loss functions. Next, we detail each component in the following sections.

3.2 Feature Projectors

The feature projectors, f_Φ and f_Ψ, map the original image or text to the corresponding high-dimensional common feature vectors. In our work, we treat the projection as a two-stage process. That is we first map the original inputs to domain-specific representations, and then use extra convolutional layers to learn the common representation. Formally, given an input image $\mathbf{I} \in \mathbb{R}^{H\times W\times 3}$, we use a backbone network (*e.g.,* VGG-16 [32]) parameterized by Φ_1 to obtain its corresponding domain-specific representation: $u = f_{\Phi_1}(\mathbf{I})$. Similarly, the given text data $\mathbf{T}$ is converted to a real-valued vector v as the domain-specific features: $v = f_{\Psi_1}(\mathbf{T})$, where the mapping function $f_{\Psi_1}(\mathbf{T})$ can be either a Text-CNN [19, 48] or Bag of Words (BoW). After that, we utilize mapping functions f_{Φ_2} and f_{Ψ_2} to further map u and v to the common representations U and V:

$$U = f_{\Phi_2}(u),\ V = f_{\Psi_2}(v), \tag{1}$$

where Φ_2 and Ψ_2 are learnable parameters. We expect that the common representations U and V should be modality-invariant and semantically discriminative.

3.3 Graph-constrained Structure Encoding

The goal of our Graph-constrained Structure Encoding (GSE) module f_χ is to fully encode structural semantics across each individual sample for producing pseudo common embeddings which will be used as semantic constraints for learning the parameters of feature projectors.

Graph Construction. The given training set $\mathcal{D}$ contains more structural information beyond independent label for each individual sample, however, they are largely ignored by existing approaches. To mine the valuable, structural semantics from the training set, we build a graph for each modality to characterize the hidden relations. Formally, we construct the fully connected graph for each modality, *i.e.,* $\mathcal{G}_u = \{\mathcal{V}_u, \mathcal{E}_u\}$ or $\mathcal{G}_v = \{\mathcal{V}_v, \mathcal{E}_v\}$, to link individual sample of data based on manual labels. That is, we treat each sample as a node and the link reflects their relations, *i.e.,* if they contain the same semantics. Given k input features, $\{u_1, \ldots, u_k\}$ or $\{v_1, \ldots, v_k\}$, within each training batch, we directly treat them as nodes of the underlying graphs $\mathcal{V}_u = \{u_1, \ldots, u_k\}$ or $\mathcal{V}_v\{v_1, \ldots, v_k\}$. If features share the same manual labels, we add a link between them $e_u(u_i, u_j) = 1$ (or $e_v(v_i, v_j) = 1$), otherwise $e_u(u_i, u_j) = 0$ (or $e_v(v_i, v_j) = 0$). Besides, the self-connections at each node are also included $e_u(u_i, u_i) = 1$ (or $e_v(v_i, v_i) = 1$).

Graph-based Structural Information Encoding. The core idea of our GSE is to perform K message propagation iterations over $\mathcal{G}_u$ or $\mathcal{G}_v$ to efficiently encode rich and intra-modality structural information within a group of training samples. This helps to better capture the high-level semantic information from a global view and obtain discriminative representation embeddings.

Our GSE consists of a two-branch graph convolutional network (GCN) to achieve above goal, in which each graph branch reasons over samples for one modality by following the standard layer-wise propagation rule:

$$\begin{aligned} \mathcal{V}_u{}^{(l+1)} &= \sigma\left(\widetilde{D}^{-\frac{1}{2}}\mathcal{E}_u\widetilde{D}^{-\frac{1}{2}}\mathcal{V}_u{}^{(l)}W_u^{(l)}\right), \\ \mathcal{V}_v{}^{(l+1)} &= \sigma\left(\widetilde{D}^{-\frac{1}{2}}\mathcal{E}_v\widetilde{D}^{-\frac{1}{2}}\mathcal{V}_v{}^{(l)}W_v^{(l)}\right), \end{aligned} \tag{2}$$

where $\mathcal{E}_u$ and $\mathcal{E}_v$ are the adjacency (edge) matrixs of $\mathcal{G}_u$ and $\mathcal{G}_v$ respectively, and $\mathcal{V}_u^{(l)}$ and $\mathcal{V}_v^{(l)}$ mean the node representations for images and texts at the l^{th} layer. $\widetilde{D}$ denotes the degree matrix, $W_u^{(l)}$ and $W_v^{(l)}$ are the weight matrices of l^{th} layer, and $\sigma(\cdot)$ stands for a

non-linear activation function. After L layers, the pseudo common representations can be achieved: $\check{\mathbf{U}} = \mathcal{V}_u{}^{(L)}$ and $\check{\mathbf{V}} = \mathcal{V}_u{}^{(L)}$.

Auxiliary Classifier and Loss Function. To enforce the graph-reasoned pseudo representations ($\check{\mathbf{U}}$ and $\check{\mathbf{V}}$) to be both modality-invariant and semantically discriminative, we employ a shared classifier f_{class} as the auxiliary module. Our idea is to feed both $\check{\mathbf{U}}$ and $\check{\mathbf{V}}$ to f_{class} and use the Frobenius norm to re-formulate the widely-used discrimination loss:

$$\mathcal{L}_{class} = \frac{1}{k}\left\|f_{class}(\check{\mathbf{U}}) - Y\right\|_F + \frac{1}{k}\left\|f_{class}(\check{\mathbf{V}}) - Y\right\|_F, \quad (3)$$

where $\|\cdot\|_F$ means the Frobenius norm, Y means the classification label and k means the number of samples. Eq. 3 encourages data with the same semantic label to have the same underlying representations, thus it increases the discriminative power while reduces the cross-modality discrepancy.

Moreover, we also directly minimize the distance between features of different modalities if they belong to the same category. The distance loss $\mathcal{L}_{dis}$ is defined as:

$$\mathcal{L}_{dis} = \frac{1}{k}\left\|\check{\mathbf{U}} - \check{\mathbf{V}}\right\|_F, \quad (4)$$

Therefore, the total objective function $\mathcal{L}_{gse}$ for training our GSE is a combination of Eq. 3 and Eq. 4:

$$\mathcal{L}_{gse} = \mathcal{L}_{class} + \lambda \cdot \mathcal{L}_{dis}, \quad (5)$$

where λ is the combination weight.

Our GSE can be plugged into conventional common feature projectors to encode graph-structured semantics and create pseudo common representations that can be acted as extra supervisions. It also should be mentioned that GSE will be removed at the inference stage, so it introduces no additional computation cost to the common feature projectors.

3.4 GAN-based Dual Learning Network

We further upgrade our projectors by using the adversarial learning strategy, such that the joint distribution over the data across modalities can be better modelled. Specifically, as shown in Figure 2, our GAN-based Dual Learning (GDL) network unites (i) GAN-based Image Feature Projector and (ii) GAN-based Text Feature Projector with two discriminative models D_δ and D_ρ. It aims at further pushing the embedding space of image and text together to enhance the cross-modality invariant.

GAN-based Image Feature Projector. Our GAN-based image feature projector includes a generative model and a discriminative model. It takes the common image feature projector as the generative model which transfers domain-specific image features to the common representation by $f_{\Phi_2}: u \rightarrow U$; Then, a newly-added discriminative model is used to estimate the probability that a sample came from $f_{\Psi_2}(v)$ rather than U by D_δ: U or $f_{\Psi_2}(v) \rightarrow \{0, 1\}$. The adversarial objective function can be formulated as follows:

$$\mathcal{L}_{adv_u} = -\frac{1}{k}\sum_{i=1}^{k}\{(\log D_\delta(f_{\Psi_2}(v_i))) + \log(1 - D_\delta(U_i))\} \quad (6)$$

where k denotes the number of samples.

GAN-based Text Feature Projector. Similar to the GAN-based Image Feature Projector, the generative model in the GAN-based Text Feature Projector maps the domain-specific text features to the common representation by $f_{\Psi_2}: v \rightarrow V$. The discriminative model D_ρ treats the text features V generated by f_{Ψ_2} as fake data and features U from f_{Φ_2} as the real data, and tries to classify between the fake and real data. Thus, the loss function is defined as:

$$\mathcal{L}_{adv_v} = -\frac{1}{k}\sum_{i=1}^{k}\{(\log D_\rho(f_{\Phi_2}(u_i))) + \log(1 - D_\rho(V_i))\} \quad (7)$$

As the above two GAN-based projectors are naturally mixed or interwoven together by our GDL, the inter-modality information can be easily explored without using any manual labels.

Algorithm 1: Symmetrical Training Algorithm

Initialization: Training data: $\mathcal{D} = (\mathcal{I}, \mathcal{T}, \mathcal{Y})$;
Mini-batch size: k;
maximum iteration number: $iter_{max}$, $iter = 0$.
while iter < $iter_{max}$ **do**
  Sample the images, texts and labels: $[\mathbf{I}_i, \mathbf{T}_i, \mathbf{Y}_i]_{i=1}^{k} \subseteq \mathcal{D}$;
  Training for GSE:
  Update parameters for GSE module: f_χ, f_{Φ_1} and f_{Ψ_1} by Eq. 5;
  Fix f_χ, f_{Φ_1}, f_{Ψ_1};
  Training for GDL:
  Fix f_{Φ_2}, f_{Ψ_2}, and update parameters for D_δ by Eq. 12;
  Fix D_δ, and update parameters for f_{Φ_2} by Eq. 13;
  Fix f_{Φ_2}, f_{Ψ_2}, and update parameters for D_ρ by Eq. 14;
  Fix D_ρ, and update parameters for f_{Ψ_2} by Eq. 15;
end while

3.5 Graph-constrained Cross-modal Retrieval

As shown in Figure 2, our fully-equipped Graph-constrained Cross-modal Retrieval (GCR) model integrates feature projectors, GSE and GDL within a unified learning framework.

Objective Function. We design multiple loss functions to train our GCR for producing reliable common representations, U and V, for CMR. Generally, it should be trained to map the original data of different modalities to the common feature representations which are modality-invariant and semantically discriminative. First, to capture the graph-structured knowledge learned by GSE, we design the following graph constrained loss $\mathcal{L}_{GC}$ function:

$$\mathcal{L}_{GC} = \frac{1}{k}\left\|\check{\mathbf{U}} - U\right\|_F + \frac{1}{k}\left\|\check{\mathbf{V}} - V\right\|_F, \quad (8)$$

where $\|\cdot\|_F$ is the Frobenius norm which encourages the small feature distances between U and the pseudo common features $\check{\mathbf{U}}$, as well as V and the pseudo common features $\check{\mathbf{V}}$.

Moreover, we also reuse $\mathcal{L}_{class}$ and feed U and V to the shared classifier f'_{class} to further enforce the modality-invariant representations:

$$\mathcal{L}_{class} = \frac{1}{k}\left\|f'_{class}(U) - Y\right\|_F + \frac{1}{k}\left\|f'_{class}(V) - Y\right\|_F, \quad (9)$$

where $f'_{class}(\cdot)$ means the well-trained classifier within the GSE module. Thus, the total loss $\mathcal{L}_{GSS}$ for capturing graph-structured semantics is given as:

$$\mathcal{L}_{GSS} = \mathcal{L}_{GC} + \tau \cdot \mathcal{L}_{class}, \quad (10)$$

where τ means the combination weight.

Second, the losses for GDL (Eq. 6 and Eq. 7) are also included for training our GCR. Thus, the total loss $\mathcal{L}_{GDL}$ for our GCR is a combination of all losses functions.

$$\mathcal{L}_{GDL} = \mathcal{L}_{GSS} + \mathcal{L}_{adv_u} + \mathcal{L}_{adv_v}. \tag{11}$$

Symmetrical Training Algorithm. We can minimize Eq. 11 to learn all parameters in our GCR for achieving the desired projection functions f_Φ and f_Ψ. However, as there are many learnable components for different functions, GCR can not be optimized as a whole. Here, we introduce a symmetrical training algorithm to decouple the training for the whole GCR to small sequential steps.

Specifically, we first optimize the GSE module f_χ along with f_{Φ_1} and f_{Ψ_1}, given the manual labels, by minimizing the loss function of Eq. 5. Once trained, we fix the parameters of our GSE module (f_χ, f_{Φ_1} and f_{Ψ_1}), and learn parameters of our GDL. To better learn parameters of GDL, including f_{Φ_2}, f_{Ψ_2}, D_δ and D_ρ, we draw inspiration from the expectation–maximization (EM) algorithm [7], and fix some parameters and learn the others in an interactive manner. That is, we first fix the generated features from f_{Φ_2} and f_{Ψ_2}: $U = f_{\Phi_2}(u)$ and $V = f_{\Phi_2}(v)$, and train discriminative model D_δ by re-writing Eq. 11 as,

$$\mathcal{L}_{dis_u} = -\frac{1}{k}\sum_{i=1}^{k}\{(\log D_\delta(V_i)) + \log(1 - D_\delta(U_i))\}, \tag{12}$$

After that, we fixed parameters of D_δ, and learn the generative model f_{Φ_2} by

$$\mathcal{L}_{gen_u} = -\frac{1}{k}\sum_{i=1}^{k}\{(\log D_\delta(f_{\Phi_2}(u_i)))\} + \alpha \cdot \mathcal{L}_{GSS_u}, \tag{13}$$

where $\mathcal{L}_{GSS_u} = \frac{1}{k}\left\|\check{\mathbf{U}} - U\right\|_F + \tau \cdot \frac{1}{k}\left\|f'_{class}(U) - Y\right\|_F$.

Similarly, for learning the D_ρ, we use the following loss function,

$$\mathcal{L}_{dis_v} = -\frac{1}{k}\sum_{i=1}^{k}\{(\log D_\rho(U_i)) + \log(1 - D_\rho(V_i))\}, \tag{14}$$

and then learn f_{Ψ_2} by

$$\mathcal{L}_{gen_v} = -\frac{1}{k}\sum_{i=1}^{k}\{(\log D_\rho(f_{\Psi_2}(v_i)))\} + \beta \cdot \mathcal{L}_{GSS_v}, \tag{15}$$

where $\mathcal{L}_{GSS_v} = \frac{1}{k}\left\|\check{\mathbf{V}} - V\right\|_F + \tau \cdot \frac{1}{k}\left\|f'_{class}(V) - Y\right\|_F$. More details about the symmetrical training algorithm above are given in Algorithm 1.

3.6 Implementation Details

Network Configuration. For domain-specific image feature projectors f_{Φ_1}, most backbone can be used to implement f_{Φ_1}, and we test two commonly used ones,*i.e.,* VGG-16 and VGG-19 [32]. For domain-specific text feature projectors f_{Ψ_1}, BoW is used to implement f_{Ψ_1} for fair comparison with most exiting arts. Following [17], the 3,000-dimensional, 1,000-dimensonal, and 3,000-dimensional BoW features are separately used for Wikipedia, NUS-WIDE-10K, and XMedia datasets. For the Pascal Sentence dataset, we follow [48] to learn a 300-dimensional representation vector for each text. It is also noteworthy that Text CNN [48] is also applicable for extracting domain-specific text features. Then, for f_{Φ_2} and f_{Ψ_2}, we implement them by a three-layer feed-forward neural network. As for the GSE, we use three GCN layers with ReLU for each modality to encode graph-structured information. The shared classifier f_{class} is implemented by a fully-connected layer. Both of the discriminative models D_δ and D_ρ are implemented as the three-layer feed-forward neural networks.

Training Details. We train our GCR model by using Algorithm 1. Specifically, the GCR model is trained using the Adam optimizer with a batch size of $k = 128$. We start the training with the learning rate of 10^{-4}. We set $\lambda = 0.25$ in Eq. 4, $\tau = 12$, $\alpha = 1$ in Eq. 13 and $\beta = 1.2$ in Eq. 15. A total of 200 epochs are used for training.

Inference. Once trained, the feature projectors can be directly applied to unseen samples for CMR. It also be noted that GSE (f_χ) and discriminator models (D_u and D_v) only used in the training stage, and will be removed in the inference stage.

Reproducibility. Our model is implemented on PyTorch and trained on a GeForce GTX 1080 Ti. To provide full details of our method, all codes will be released.

4 EXPERIMENTS

4.1 Experimental Settings

Dataset. Four commonly-used benchmark datasets are used for performance evaluation. Wikipedia [6] consists of 2,866 image-text pairs, which is generated from Wikipedia's "featured article". NUS-WIDE-10K [5] has 10,000 labeled images, in which each image is associated with user tags. XMedia [26, 45] is a large-scale cross-media dataset which consists of five media types. Pascal Sentence [27] contains 1,000 images, each of which is annotated with five English sentences describing its content.

Evaluation Protocol. To fairly benchmark our GCR, we follow [24, 35, 48], and adopt the mean Average Precision (mAP) as the evaluation metric to compare all CMR approaches.

4.2 Comparison with State-of-the-Arts

We compare our method with 10 state-of-the-art (SOTA) approaches, including 2 traditional cross-modal retrieval methods, namely CCA [12] and JRL [46], and 8 deep learning models, namely Corr-AE [9], DCCA [2], CMDN [23], Deep-SM [37] , DSCMR [48], ACMR [35], CM-GANs [24] and MHTN [17].

Performance on Wikipedia: Table 1 reports the comparison results with 10 SOTAs on Wikipedia. As can be seen, our GCR$_{vgg16}$ model achieves very competitive results compared to its counterparts. When compared with the state-of-the-art DSCMR$_{vgg16}$ [48] and MHTN$_{vgg16}$ [17], our GCR$_{vgg16}$ achieves clear performance improvements. Moreover, when using stronger VGG-19 as the backbone for extracting domain-specific image features, our GCR$_{vgg19}$ model further boosts the performance and sets a new record. Furthermore, if we further replace the commonly used BoW by Text CNN [47] (as in [48]) for extracting text features, the performance can be largely improved. Clearly, our solution can enable the common feature projectors to be trained for better extracting modality-invariant and discriminative features.

Performance on NUS-WIDE-10K: On the large-scale NUS-WIDE-10K, our method also achieves promising results. As shown in Table 2, our GCR$_{vgg16}$ has already achieved the average mAP score of 0.538, which is higher than most existing SOTA approaches. With

Table 1: Performance comparison in terms of mAP on the Wikipedia. '†' denotes the methods using VGG-19 for extracting domain-specific image features; '‡' means that Text CNN [47] is used to extract domain-specific text features.

Methods	Image2Text	Text2Image	Average
CCA [12]	0.176	0.178	0.177
CCA† [12]	0.258	0.250	0.254
JRL [46]	0.408	0.353	0.381
JRL‡ [46]	0.449	0.418	0.434
Corr-AE [9]	0.373	0.357	0.365
Corr-AE† [9]	0.402	0.395	0.399
DCCA [2]	0.409	0.355	0.382
DCCA‡ [2]	0.444	0.396	0.420
Deep-SM [37]	0.458	0.345	0.402
CMDN [23]	0.409	0.364	0.387
CMDN‡[23]	0.487	0.427	0.457
DSCMR [48]	0.516	0.462	0.489
DSCMR‡ [48]	0.521	0.479	0.500
ACMR [35]	0.439	0.361	0.400
ACMR† [35]	0.518	0.412	0.465
ACMR‡ [35]	0.511	0.467	0.489
MHTN [17]	0.527	0.446	0.487
MHTN† [17]	0.541	0.461	0.501
GCR (ours)	**0.538**	**0.491**	**0.515**
GCR† (ours)	**0.553**	**0.501**	**0.527**
GCR‡ (ours)	**0.584**	**0.528**	**0.556**

Table 2: Performance comparison in terms of mAP scores on the NUSWIDE-10K. '†' denotes the approaches using VGG-19 for extracting domain-specific image features.

Methods	Image2Text	Text2Image	Average
CCA [12]	0.159	0.189	0.174
CCA† [12]	0.258	0.250	0.254
Corr-AE [9]	0.306	0.340	0.323
Corr-AE† [9]	0.402	0.395	0.399
Deep-SM [37]	0.389	0.496	0.443
CMDN [23]	0.410	0.450	0.430
CMDN† [23]	0.492	0.515	0.504
DSCMR [48]	0.512	0.527	0.520
DSCMR† [48]	0.552	0.542	0.547
ACMR [35]	0.445	0.473	0.459
ACMR†[35]	0.544	0.538	0.541
MHTN [17]	0.520	0.534	0.527
MHTN†[17]	0.552	0.541	0.547
GCR (ours)	**0.535**	**0.540**	**0.538**
GCR (ours) †	**0.554**	**0.548**	**0.551**

a stronger VGG-19, our GCR_{vgg19} set a new record on NUS-WIDE-10K. This is because our approach can fully mine graph-structured semantics from the training set to guide the learning of common feature projectors, which makes the common feature projectors learn to extract features from a global view.

Performance on XMedia: We also compare our GCR with SOTAs on XMedia. Again, our approach achieves the best performance as shown in Table 3. The powerful graph-based GSE module and GAN-based GDL enable the common feature projectors to be well trained.

Table 3: Performance comparison in terms of mAP scores on the XMedia dataset.

Methods	Image2Text	Text2Image	Average
CCA [12]	0.257	0.341	0.299
JRL [46]	0.770	0.800	0.785
Corr-AE [9]	0.450	0.437	0.443
DCCA [2]	0.472	0.466	0.469
Deep-SM [37]	0.822	0.807	0.814
CMDN [23]	0.794	0.805	0.799
DSCMR [48]	0.865	0.853	0.859
ACMR [35]	0.704	0.710	0.707
MHTN [17]	0.853	0.843	0.848
GCR (ours)	**0.892**	**0.879**	**0.886**

Table 4: Performance comparison in terms of mAP scores on the Pascal Sentence dataset.

Methods	Image2Text	Text2Image	Average
CCA [12]	0.225	0.227	0.226
JRL [46]	0.527	0.534	0.531
Corr-AE [9]	0.532	0.521	0.527
DCCA [2]	0.678	0.677	0.678
Deep-SM [37]	0.560	0.539	0.550
CMDN [23]	0.544	0.526	0.535
DSCMR [48]	0.710	0.722	0.716
ACMR [35]	0.671	0.676	0.673
CM-GANs [24]	0.603	0.604	0.604
GCR (Ours)	**0.726**	**0.730**	**0.728**

Thus, once trained, they can more reliably project the raw image or text to the common feature space for retrieval across modalities.

Performance Pascal Sentence: The quantitative comparison with 9 SOTA approaches on Pascal Sentence is summarized in Table 4. We find that our GCR again achieves the best performance according to mAP score. The high accuracy should be attributed to the graph-constrained common embedding learning approach. It is also noteworthy that our projectors require no additional computational and memory costs during the inference stage.

4.3 Ablation Study

In this section, we study the efficacy of our core ideas and essential model designs, over Wikipedia [6]. We adopt VGG-16 [32] as the domain-specific image feature projector, and BoW as domain-specific text features. To perform extensive ablation experiments, we train each model for 200 epochs while keeping other hyper-parameters unchanged.

To verify the effectiveness of each novel module, we provide multiple baseline models for comparison. The baseline model is designed by removing our GSE and GDL modules, and utilizes the text projectors to constrain the training of image projectors, and vice versa. With our GDL, as shown in Table 5, we observe clear performance improvements. In addition, we also provide a model by using the conventional GAN techniques to train each feature projector independently (Baseline + GAN). Clearly, our GDL is superior to the conventional strategy, and achieves higher mAP scores. Finally, with our GSE, the fully-equipped approach is able to obtain the highest mAP scores.

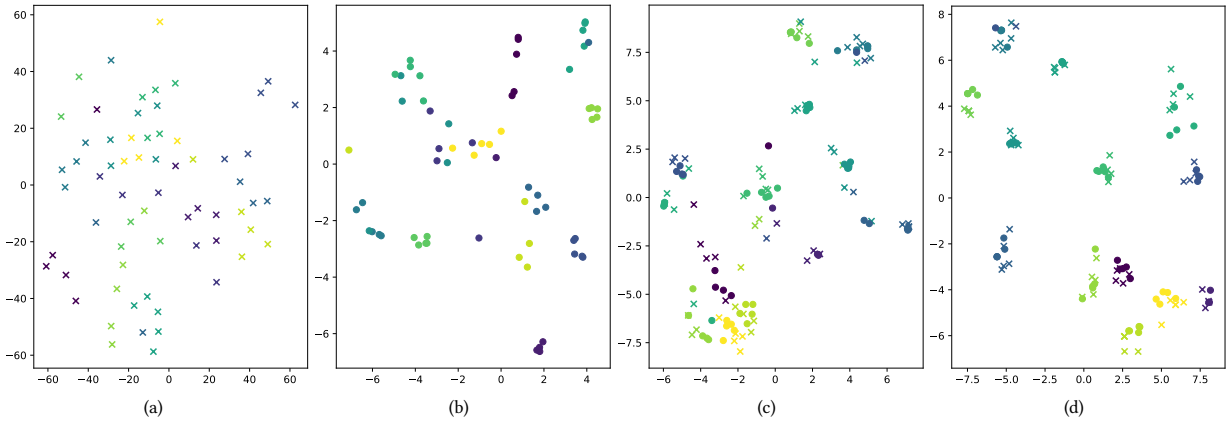

Figure 3: The visualisation for the test data on Pascal Sentence [27] by using the t-SNE method [34]. (a) the original image representations. (b) the original text representations. (c) the image and text representations in the common representation space after using GDL. (d) the image and text representations in the common representation space after using GDL and GSE.

Table 5: The ablation results of our proposed approach in terms of mAP scores on the Wikipedia dataset.

Methods	Image2Text	Text2Image	Average
Baseline	0.413	0.421	0.417
Baseline + GAN	0.491	0.447	0.469
Baseline + GDL	0.513	0.462	0.488
Baseline + GDL + GSE (ours)	**0.538**	**0.491**	**0.515**

4.4 Visualization

To further study the effectiveness of our novel designs, we use t-SNE tool to visualize the distribution of the generated features on Pascal Sentence [27]. We randomly select 13 from 20 categories to visualize the representations in the common representation space. As can be seen in Figure 3 (a) and (b), the original representations for both texts and images are neither discriminative nor compact. With our GDL, the projectors can be well trained, thus they produce more consistent representations across modalities, as shown in Figure 3 (c). However, some categories are still close to each other and not well separated. Finally, with our GSE, the learned common representations become more compact and well separated. This suggests that, by exploring the graph-structured semantics across individual samples can ensure the projectors to generate more modality-invariant and discriminative features, hence achieving promising results for cross-modal retrieval.

4.5 Convergency

We also provide the convergence curves of all losses on XMedia dataset, as shown in Figure 4. We can observe that all the five losses decrease almost monotonously and converges smoothly, and become stable after 100 epochs. Thus, our model is easy to train and can converge well and fast. Because our GSE is used to guide the training of generative models to better capture high-level semantics, the losses for generative models seem slightly undulant. Especially, the losses of discriminative model experience slight fluctuation in the initial few epochs and then stabilizes. These results are in accordance with our expectation, since the goal of adversarial learning is to make the discriminative model unable to distinguish real data from fake data.

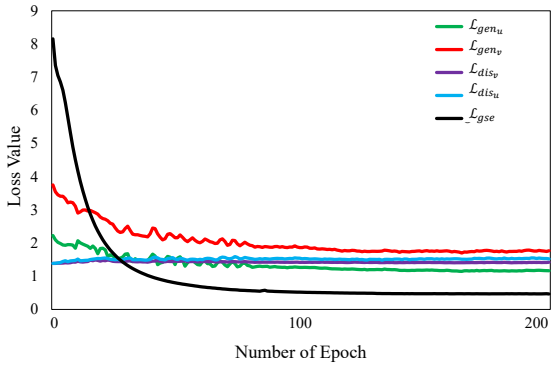

Figure 4: Convergence curves of XMedia dataset.

5 CONCLUSION

In this paper, we proposed a novel approach for cross-modal retrieval, called Graph-constrained Cross-modal Retrieval (GCR), to fully explore graph-structured semantics within each training batch for cross-modal retrieval. A novel Graph-constrained Structure Encoding (GSE) is introduced to mine and encode the structural information from training set, and produce pseudo common representations that can be acted as extra supervisions. The underlying GAN-based Dual Learning network (GDL) further push the embedding space of image and text together to enhance the cross-modality invariant. Extensive experiments show that explicitly mining structural information from the dataset can help to learn modality-invariant and discriminative representations, hence improving the accuracy of cross-modal retrieval. We believe our novel designs, GSE and GDL can also benefit other related multimedia tasks, such as audio-visual learning.

ACKNOWLEDGMENTS

This work was supported by Key-Area Research and Development Program of Guangdong Province (No.2019B010136003), Sichuan Science and Technology Program (No.2019YJ0176, No.2019YJ0177, No.2019YFQ0005).

REFERENCES

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Salt Lake City, UT, USA, 6077–6086. https://doi.org/10.1109/CVPR.2018.00636

[2] Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. 2013. Deep Canonical Correlation Analysis. In *Proceedings of the 30th International Conference on International Conference on Machine Learning* (Atlanta, GA, USA) *(ICML'13, Vol. 28)*. JMLR.org, Atlanta, GA, USA, 1247–1255.

[3] Joan Bruna, Wojciech Zaremba, Arthur Szlam, and Yann LeCun. 2014. Spectral Networks and Locally Connected Networks on Graphs. arXiv:1312.6203 [cs.LG]

[4] Yongming Chen, Liang Wang, Wei Wang, and Zhang Zhang. 2012. Continuum regression for cross-modal multimedia retrieval. In *2012 19th IEEE International Conference on Image Processing*. IEEE, Orlando, FL, USA, 1949–1952. https://doi.org/10.1109/ICIP.2012.6467268

[5] Tatseng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng. 2009. NUS-WIDE: A Real-world Web Image Database from National University of Singapore. In *in: Proceedings of the ACM International Conference on Image and Video Retrieval* (Santorini, Fira, Greece) *(CIVR '09)*. ACM, New York, NY, USA, Article 48, 9 pages.

[6] Jose Costa Pereira, Emanuele Coviello, Gabriel Doyle, Nikhil Rasiwasia, Gert R.G. Lanckriet, Roger Levy, and Nuno Vasconcelos. 2014. On the Role of Correlation and Abstraction in Cross-Modal Multimedia Retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36, 3 (2014), 521–535. https://doi.org/10.1109/TPAMI.2013.142

[7] Arthur Dempster, Natalie Laird, and Donald Rubin. 1977. Maximum Likelihood From Incomplete Data Via The EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)* 39 (01 1977), 1–38.

[8] Chaorui Deng, Qi Wu, Qingyao Wu, Fuyuan Hu, Fan Lyu, and Mingkui Tan. 2018. Visual Grounding via Accumulated Attention. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, Salt Lake City, UT, USA, 7746–7755. https://doi.org/10.1109/CVPR.2018.00808

[9] Fangxiang Feng, Xiaojie Wang, and Ruifan Li. 2014. Cross-Modal Retrieval with Correspondence Autoencoder. In *Proceedings of the 22nd ACM International Conference on Multimedia* (Orlando, Florida, USA) *(MM '14)*. Association for Computing Machinery, New York, NY, USA, 7–16. https://doi.org/10.1145/2647868.2654902

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

[11] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative Adversarial Nets. In *Proceedings of the 27th International Conference on Neural Information Processing Systems* (Montreal, Canada) *(NIPS'14, Vol. 2)*. MIT Press, Cambridge, MA, USA, 2672–2680.

[12] David Hardoon, Sandor Szedmak, and John Shawe-Taylor. 2004. Canonical Correlation Analysis: An Overview with Application to Learning Methods. *in: Neural computation* 16 (2004), pp. 2639–2664.

[13] Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tieyan Liu, and Wei-Ying Ma. 2016. Dual Learning for Machine Translation. In *Advances in Neural Information Processing Systems 29* (advances in neural information processing systems 29 ed.). Curran Associates, Inc., Centre Convencions Internacional Barcelona, Barcelona, SPAIN, 820–828.

[14] Harold Hotelling. 1935. Relations Between Two Sets of Variates. *Biometrika* 28 (11 1935), 321–377. https://doi.org/10.1093/biomet/28.3-4.321

[15] Peng Hu, Liangli Zhen, Dezhong Peng, and Pei Liu. 2019. Scalable Deep Multimodal Learning for Cross-Modal Retrieval. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Paris, France) *(SIGIR'19)*. Association for Computing Machinery, New York, NY, USA, 635–644. https://doi.org/10.1145/3331184.3331213

[16] Ronghang Hu, Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Kate Saenko. 2017. Learning to Reason: End-to-End Module Networks for Visual Question Answering. In *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, Venice, Italy, 804–813. https://doi.org/10.1109/ICCV.2017.93

[17] Xin Huang, Yuxin Peng, and Mingkuan Yuan. 2020. MHTN: Modal-Adversarial Hybrid Transfer Network for Cross-Modal Retrieval. *IEEE Transactions on Cybernetics* 50, 3 (2020), 1047–1059. https://doi.org/10.1109/TCYB.2018.2879846

[18] Yash Khare, Viraj Bagal, Minesh Mathew, Adithi Devi, U Priyakumar, and CV Jawahar. 2021. MMBERT: Multimodal BERT Pretraining for Improved Medical VQA. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*. IEEE, Nice, France, 1033–1036. https://doi.org/10.1109/ISBI48211.2021.9434063

[19] Yoon Kim. 2014. Convolutional Neural Networks for Sentence Classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Doha, Qatar, 1746–1751. https://doi.org/10.3115/v1/D14-1181

[20] Ao Luo, Xin Li, Fan Yang, Zhicheng Jiao, Hong Cheng, and Siwei Lyu. 2020. Cascade graph neural networks for rgb-d salient object detection. In *European Conference on Computer Vision*. Springer, 346–364.

[21] Ao Luo, Fan Yang, Xin Li, Dong Nie, Zhicheng Jiao, Shangchen Zhou, and Hong Cheng. 2020. Hybrid graph neural networks for crowd counting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 11693–11700.

[22] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Ng. 2011. Multimodal Deep Learning. In *Proceedings of the 28th International Conference on Machine Learning (ICML 2011)*. Omnipress, Bellevue, Washington, USA, 689–696.

[23] Yuxin Peng, Xin Huang, and Jinwei Qi. 2016. Cross-Media Shared Representation by Hierarchical Learning with Multiple Deep Networks. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI'16)*. AAAI Press, New York, New York, USA, 3846–3853.

[24] Yuxin Peng and Jinwei Qi. 2019. CM-GANs: Cross-Modal Generative Adversarial Networks for Common Representation Learning. *ACM Trans. Multimedia Comput. Commun. Appl.* 15, 1, Article 22 (Feb. 2019), 24 pages. https://doi.org/10.1145/3284750

[25] Yuxin Peng, Jinwei Qi, Xin Huang, and Yuxin Yuan. 2018. CCL: Cross-modal Correlation Learning With Multigrained Fusion by Hierarchical Network. *in: IEEE Transactions on Multimedia* 20, 2 (Feb 2018), pp. 405–420.

[26] Yuxin Peng, Xiaohua Zhai, Yunzhen Zhao, and Xin Huang. 2016. Semi-Supervised Cross-Media Feature Learning With Unified Patch Graph Regularization. *IEEE Transactions on Circuits and Systems for Video Technology* 26, 3 (2016), 583–596.

[27] Cyrus Rashtchian, Peter Young, Micah Hodosh, and Julia Hockenmaier. 2010. Collecting Image Annotations Using Amazon's Mechanical Turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk* (Los Angeles, California, USA) *(CSLDAMT '10)*. Association for Computational Linguistics, Los Angeles, California, USA, 139–147.

[28] Nikhil Rasiwasia, Jose Costa Pereira, Emanuele Coviello, Gabriel Doyle, Gert R.G. Lanckriet, Roger Levy, and Nuno Vasconcelos. 2010. A New Approach to Cross-Modal Multimedia Retrieval. In *Proceedings of the 18th ACM International Conference on Multimedia* (Firenze, Italy) *(MM '10)*. Association for Computing Machinery, New York, NY, USA, 251–260. https://doi.org/10.1145/1873951.1873987

[29] Fawaz Sammani and Luke Melas-Kyriazi. 2020. Show, Edit and Tell: A Framework for Editing Image Captions. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Seattle, WA, USA, 4807–4815. https://doi.org/10.1109/CVPR42600.2020.00486

[30] Abhishek Sharma and David Jacobs. 2011. Bypassing synthesis: PLS for face recognition with pose, low-resolution and sketch, In 2011 Conference on Computer Vision and Pattern Recognition (CVPR). *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 1, 593 – 600. https://doi.org/10.1109/CVPR.2011.5995350

[31] Abhishek Sharma, Abhishek Kumar, H. Daume, and D.W. Jacobs. 2012. Generalized Multiview Analysis: A discriminative latent space. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, Providence, RI, USA, 2160–2167. https://doi.org/10.1109/CVPR.2012.6247923

[32] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition.

[33] Joshua B Tenenbaum and William T Freeman. 2000. Separating style and content with bilinear models. *Neural computation* 12, 6 (2000), 1247–1283.

[34] Laurens van der Maaten and Geoffrey Hinton. 2008. Viualizing data using t-SNE. *Journal of Machine Learning Research* 9 (11 2008), 2579–2605.

[35] Bokun Wang, Yang Yang, Xing Xu, Alan Hanjalic, and Heng Tao Shen. 2017. Adversarial Cross-Modal Retrieval. In *Proceedings of the 25th ACM International Conference on Multimedia* (Mountain View, California, USA) *(MM '17)*. Association for Computing Machinery, New York, NY, USA, 154–162. https://doi.org/10.1145/3123266.3123326

[36] Xin Wei, Ruixuan Yu, and Jian Sun. 2020. View-GCN: View-Based Graph Convolutional Network for 3D Shape Analysis. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Seattle, WA, USA, 1847–1856. https://doi.org/10.1109/CVPR42600.2020.00192

[37] Yunchao Wei, Yao Zhao, Canyi Lu, Shikui Wei, Luoqi Liu, Zhenfeng Zhu, and Shuicheng Yan. 2017. Cross-Modal Retrieval With CNN Visual Features: A New Baseline. *IEEE Transactions on Cybernetics* 47, 2 (2017), 449–460. https://doi.org/10.1109/TCYB.2016.2519449

[38] Yuan Xie, Tianshui Chen, Tao Pu, Hefeng Wu, and Liang Lin. 2020. Adversarial Graph Representation Adaptation for Cross-Domain Facial Expression Recognition. In *Proceedings of the 28th ACM International Conference on Multimedia*. Association for Computing Machinery, New York, NY, USA, 1255–1264. https://doi.org/10.1145/3394171.3413822

[39] Ruiqing Xu, Chao Li, Junchi Yan, Cheng Deng, and Xianglong Liu. 2019. Graph Convolutional Network Hashing for Cross-Modal Retrieval. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*. International Joint Conferences on Artificial Intelligence Organization, Macao, China, 982–988. https://doi.org/10.24963/ijcai.2019/138

[40] Fei Yan and Krystian Mikolajczyk. 2015. Deep correlation for matching images and text. In *2015 IEEE Conference on Computer Vision and Pattern Recognition*

(CVPR). IEEE, Boston, MA, USA, 3441–3450. https://doi.org/10.1109/CVPR.2015.7298966

[41] Han Yang, Xingjian Zhen, Ying Chi, Lei Zhang, and Xian-Sheng Hua. 2020. CPR-GCN: Conditional Partial-Residual Graph Convolutional Network in Automated Anatomical Labeling of Coronary Arteries. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Seattle, WA, USA, 3802–3810. https://doi.org/10.1109/CVPR42600.2020.00386

[42] Zili Yi, Hao Zhang, Ping Tan, and Minglun Gong. 2017. DualGAN: Unsupervised Dual Learning for Image-to-Image Translation. In *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, Venice, Italy, 2868–2876. https://doi.org/10.1109/ICCV.2017.310

[43] Jing Yu, Yuhang Lu, Zengchang Qin, Weifeng Zhang, Yanbing Liu, Jianlong Tan, and Li Guo. 2018. Modeling Text with Graph Convolutional Network for Cross-Modal Information Retrieval. In *Advances in Multimedia Information Processing – PCM 2018*. Springer International Publishing, Cham, 223–234. https://doi.org/10.1007/978-3-030-00776-8_21

[44] Qiang Zhai, Xin Li, Fan Yang, Chenglizhao Chen, Hong Cheng, and Deng-Ping Fan. 2021. Mutual Graph Learning for Camouflaged Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 12997–13007.

[45] Xiaohua Zhai, Yuxin Peng, and Jianguo Xiao. 2014. Learning Cross-Media Joint Representation With Sparse and Semisupervised Regularization. *IEEE Transactions on Circuits and Systems for Video Technology* 24, 6 (2014), 965–978.

[46] Xiaohua Zhai, Yuxin Peng, and Jianguo Xiao. 2014. Learning Cross-Media Joint Representation With Sparse and Semisupervised Regularization. *IEEE Transactions on Circuits and Systems for Video Technology* 24, 6 (2014), 965–978. https://doi.org/10.1109/TCSVT.2013.2276704

[47] Ye Zhang and Byron Wallace. 2015. A Sensitivity Analysis of (and Practitioners' Guide to) Convolutional Neural Networks for Sentence Classification.

[48] Liangli Zhen, Peng Hu, Xu Wang, and Dezhong Peng. 2019. Deep Supervised Cross-Modal Retrieval. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Long Beach, CA, USA. https://doi.org/10.1109/CVPR.2019.01064

[49] Yangchun Zhu, Zheng-Jun Zha, Tianzhu Zhang, Jiawei Liu, and Jiebo Luo. 2020. A Structured Graph Attention Network for Vehicle Re-Identification. In *Proceedings of the 28th ACM International Conference on Multimedia*. Association for Computing Machinery, New York, NY, USA, 646–654. https://doi.org/10.1145/3394171.3413607