Fine-Grained Image-Text Retrieval via Discriminative Latent Space Learning

Min Zheng, Wen Wang, and Qingyong Li

***Abstract*—Fine-grained image-text retrieval aims at searching relevant images among fine-grained classes given a text query or in a reverse way. The challenges are not only bridging the gap between two heterogeneous modalities but also dealing with large inter-class similarity and intra-class variance existed in fine-grained data. To deal with the above challenges, we propose a Discriminative Latent Space Learning (DLSL) method for fine-grained image-text retrieval. Concretely, image and text features are extracted for capturing the subtle difference in fine-grained data. Subsequently, based on the extracted features, we perform couple dictionary learning to align the heterogeneous data in a uniform latent space. To make such alignment discriminative enough for the fine-grained task, the learned latent space is endowed with discriminative property via learning a discriminative map. Comprehensive experiments on fine-grained datasets demonstrate the effectiveness of our approach.**

***Index Terms*—Discriminative latent space learning, fine-grained, image-text alignment.**

I. Introduction

With the development of multimedia technologies, multimedia data from different modalities, such as image, video, text, is utilized together to perceive real-world objects, and therefore the cross-media retrieval has been attracting more and more attention. Specifically, image-text retrieval aims to take one type (e.g., image) of data as the query to retrieve relevant data of another type (e.g., text) [1], [2]. In real-life scenarios, there are two types of image-text retrieval tasks [3], that is coarse-grained and fine-grained. The former returns results from the same category, and the latter returns results that belong to the same subcategory.

Benefit from the breakthrough of deep learning [4]–[7] and information retrieval [8]–[11], coarse-grained image-text retrieval has achieved great progress in recent years [12]–[20]. These methods usually utilize an image encoder (e.g. CNN) and a text encoder (e.g. RNN) to extract the global features of images and texts, and then devise a metric to measure the similarity of image-text pairs. Beyond these coarse-grained methods, fine-grained image-text retrieval faces the challenge of highly similar global geometry and appearance among fine-grained classes.

Manuscript received December 30, 2020; revised February 17, 2021; accepted March 1, 2021. Date of publication March 11, 2021; date of current version April 16, 2021. This work was supported in part by the the National Natural Science Foundation of China under Grant U2034211, 62006017, in part by the Fundamental Research Funds for the Central Universities under Grant 2020JBZD010. *(Corresponding author: Qingyong Li.)*

The authors are with the Beijing Key Lab of Traffic Data Analysis and Mining, Beijing Jiaotong University, Beijing 100044, China (e-mail: 16112080@bjtu.edu.cn; wen.wang@vipl.ict.ac.cn; liqy@bjtu.edu.cn).

Digital Object Identifier 10.1109/LSP.2021.3065595

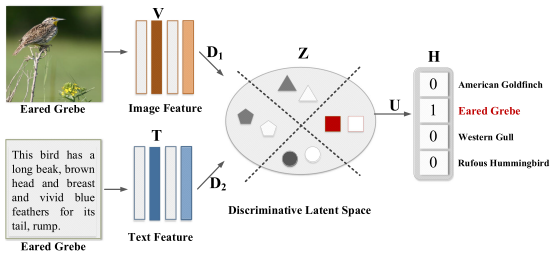

Fig. 1. Schema of DLSL. DLSL utilizes couple dictionary learning to learn a uniform latent space for the alignment of image-text pairs, and adds a discriminative property in the latent space to identify different fine-grained classes. Legends (i.e., solid and hollow) denote different modalities and polygons represent different fine-grained classes, where the red one is the exact class.

FGCross-Net [3] is firstly proposed for fine-grained representation learning based on a deep model. However, fine-grained image-text retrieval is still facing the challenge of learning the discriminative representation.

To solve this challenge, we focus on the task of fine-grained image-text retrieval and propose a novel method, named Discriminative Latent Space Learning (DLSL). Specifically, we first extract image and text representations. Then in light of the empirical success in utilizing dictionary learning to capture and correlate heterogeneous data from two different modalities [12]–[17], we propose to perform couple dictionary learning to learn a uniform latent space to achieve the alignment of heterogeneous data. Further, to enhance the discriminative ability for the subtle differences among fine-grained classes, we explore the discriminative property by learning a discriminative map for the latent space. The overall schema is shown in Fig. 1. The main contributions are summarized as follows:

- For fine-grained image-text retrieval, we propose a simple yet effective DLSL method to directly learn a common latent space by couple dictionary learning to align heterogeneous data.
- To discriminate the subtle differences among the fine-grained classes, we integrate the discriminative property into the representation of latent space.
- We divide the fine-grained image-text retrieval into two tasks: subcategory-specific and instance-specific. Experiments on both tasks demonstrate the effectiveness.

II. Discriminative Latent Space Learning

In this section, we introduce the proposed DLSL. We start with the overview, and then elaborate the details. Moreover, an effective optimization algorithm is presented. Finally, we describe the retrieval process.

A. Overview

DLSL firstly encodes images and texts by ResNet50 deep feature extractor [7] and Bag-of-words (BOW) model [21], respectively. Then in order to achieve the alignment of heterogeneous image-text pairs, DLSL constructs a uniform latent space by couple dictionary learning. Moreover, to make the latent space discriminative enough for the fine-grained task, DLSL designs a discriminative property in the latent space to identify different fine-grained classes. Finally, DLSL returns the candidates by measuring the relevance between image-text pairs in the latent space.

Formally, let $V = \{v_i\}_{i=1}^n$ be a set of image features, where v_i denotes the feature of the i-th image. Similarly, let $T = \{t_i\}_{i=1}^n$ be a set of text features, where t_i indicates the text feature of the i-th image. The goal is to learn a uniform representation $Z = \{z_i\}_{i=1}^n$, where z_i represents the uniform feature of the i-th image. Then we can perform the fine-grained image-text retrieval by measuring the similarity between the representations learned from pairs of image-text.

B. Method

Image-Text Alignment. Since images and texts are from different modalities which possess different semantics, the key issue of achieving image-text retrieval is alignment. In this work, we align the features from different modalities into a common latent feature space, so that the features from different modalities are comparable. Specifically, we present to learn a couple of dictionaries, where dictionary bases are learned in the image space and text space, respectively. Thus the corresponding reconstruction coefficients are deemed as the common latent features. Mathematically, the objective function can be formulated as:

$$\begin{aligned} \min_{D_1, D_2, Z} \ & \|V - D_1 Z\|_F^2 + \alpha \|T - D_2 Z\|_F^2, \\ \text{s.t.} \ & ||d_1^i||_2^2 \leq 1, \quad ||d_2^i||_2^2 \leq 1, \forall i, \end{aligned} \tag{1}$$

where $V \in \mathbb{R}^{j \times n}$ and $T \in \mathbb{R}^{p \times n}$ are the set of image features and text features, j, p are the dimensions of the image feature and the text feature and n is the number of samples. $D_1 \in \mathbb{R}^{j \times k}$ and $D_2 \in \mathbb{R}^{p \times k}$ are the dictionaries in the image space and the text space, where k is the dimension of the latent space. $Z \in \mathbb{R}^{k \times n}$ is the common representations of V and T in the latent space. By forcing the latent features of corresponding V and T to be same as Z, the two spaces are aligned. α is the parameter controlling the relative importance of the image space and text space. d_1^i is the i-th column of D_1 and d_2^i is the i-th column of D_2.

Discriminative Property. Additionally, in order to adapt to the fine-grained task, the representations in the latent space should be discriminative enough. That is to say, images or texts from different classes should be much more different in the latent space. Therefore, we formulate the discriminative property explicitly to classify different fine-grained classes in the latent space. The objective function can be formulated as:

$$\begin{aligned} \min_{D_1, D_2, Z, H, U} \ & \|V - D_1 Z\|_F^2 + \alpha \|T - D_2 Z\|_F^2 \\ & + \beta \|H - UZ\|_F^2 \\ \text{s.t.} \quad & ||d_1^i||_2^2 \leq 1, \quad ||d_2^i||_2^2 \leq 1, \quad ||u_i||_2^2 \leq 1, \forall i \end{aligned} \tag{2}$$

Algorithm 1: Couple Dictionary Alignment.

Input: Image features (V), Text features (T), Label matrix (H).
Output: Image dictionary (D_1), Text dictionary(D_2), Discriminative projection matrix (U) and Latent features (Z).
1: Initialize D_1, D_2, U randomly.
2: **while** *not converge* **do**
3: Update Z by (3);
4: Update D_1 by (5);
5: Update D_2 by (7);
6: Update U by (9).
7: **end while**

where $H = [h_1, h_2, \ldots, h_n] \in \mathbb{R}^{c_s \times n}$ represents the class labels of samples and c_s is the class numbers. $h_i = [0\,0 \ldots 1 \ldots 0\,0]^T \in \mathbb{R}^{c_s}$ is a one-hot label vector corresponding to sample x_i, where the non-zero entry indicates the class of the i-th image. U can be viewed as the weights of classifiers in the latent space. With such formulation, the third term in (2) aims to make the latent space to be discriminative enough to classify different classes. It implicitly pulls samples from the same class together and pushes those from different classes away from each other. α and β are regularization parameters balancing the impact of different terms.

Compared with the advanced framework [3] based on deep learning, DLSL employs the couple dictionary learning to align heterogeneous data into a latent feature space. Moreover, DLSL adds a discriminative projection matrix to enhance the discriminative ability of latent space. Such dictionary learning has a relatively weaker dependence on the volume of training samples than deep learning models. Compared with other established models using dictionary learning [13], [22]–[26], the objective functions similar to (1) and (2) are popularly applied, but DLSL is customized for fine-grained image-text retrieval by integrating discriminative property for the representation of latent space. Furthermore, DLSL learns the dictionaries simultaneously whereas the traditional methods learn the dictionary of each modality independently.

C. Optimization

Eq. (2) is not convex for D_1, D_2, U and Z simultaneously, but it is convex for each of them separately. Thus, we utilize the alternating optimization [27]. We initialize all variables to be solved and optimize them as follows.

(1) Fix D_1, D_2, U and update Z by (2). Forcing the derivative of (2) to be 0 and the closed-form solution for Z is

$$Z = (\tilde{D}^T \tilde{D})^{-1} \tilde{D}^T \tilde{X}, \tag{3}$$

where

$$\tilde{X} = \begin{bmatrix} V \\ \alpha T \\ \beta H \end{bmatrix}, \tilde{D} = \begin{bmatrix} D_1 \\ \alpha D_2 \\ \beta U \end{bmatrix}.$$

(2) Fix Z and update D_1. The subproblem can be formulated as:

$$\min_{D_1} \|V - D_1 Z\|_F^2 \quad \text{s.t.} \quad ||d_1^i||_2^2 \leq 1, \forall i. \tag{4}$$

TABLE I
THE STATISTICS OF DATA SOURCES

Task	Data source
Subcategory-specific	*Data source 1 + Data source 2*
Instance-specific	*Data source 1 + Data source 3*

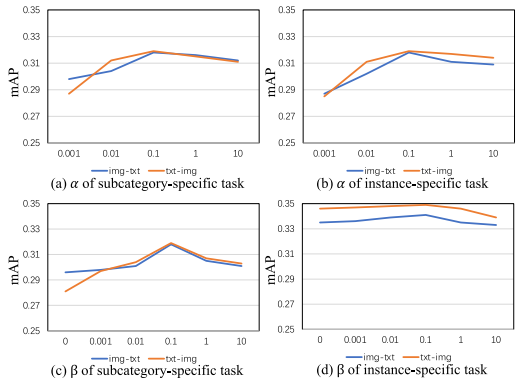

Fig. 2. Analysis of parameter α and β.

This problem can be optimized by the Lagrange dual [28]. Thus the analytical solution for (4) is

$$D_1 = (VZ^T)(ZZ^T + \Lambda)^{-1}, \tag{5}$$

where Λ is a diagonal matrix construct by all the Lagrange dual variables.

(3) Fix Z and update D_2. The subproblem can be formulated as:

$$\min_{D_2} \|T - D_2 Z\|_F^2 \quad \text{s.t.} \quad ||d_2^i||_2^2 \leq 1, \forall i. \tag{6}$$

Similarly, this problem can be also optimized by the Lagrange dual. Thus the analytical solution for (6) is

$$D_2 = (TZ^T)(ZZ^T + \Lambda)^{-1}. \tag{7}$$

(4) Fix Z and update U. The subproblem can be formulated as:

$$\min_{U} \|H - UZ\|_F^2 \quad \text{s.t.} \quad ||u_i||_2^2 \leq 1, \forall i. \tag{8}$$

Similarly,

$$U = (HZ^T)(ZZ^T + \Lambda)^{-1}. \tag{9}$$

The complete process is summarized in Algorithm 1.[1]

D. Retrieval Procedure

Considering the retrieval procedure of image-to-text (image query versus text gallery), we first extract the feature of query image by ResNet50, and project the feature into the latent feature space by

$$Z_q^* = \arg\min_{Z_q} \|V_q - D_1 Z_q\|_F^2 + \gamma \|Z_q\|_F^2, \tag{10}$$

where V_q represents the feature of query image and Z_q is the corresponding feature in the latent space. Second we learn the feature of gallery text by BOW, and project the feature into the latent feature space by

$$Z_g^* = \arg\min_{Z_g} \|T_g - D_2 Z_g\|_F^2 + \gamma \|Z_g\|_F^2, \tag{11}$$

where T_g represents the feature of gallery text and Z_g is the corresponding feature in the latent space. Then we employ cosine distance as the score function to measure the similarity between Z_q^* and Z_g^*. Finally, we return the candidate (text) based on the maximum similarity.

Likewise, the retrieval procedure for text-to-image (text query versus image gallery) can be conducted in the same way. The best value for γ is chosen by five-fold cross-validation, and the scope is set in [0.001, 0.01, 0.1, 1, 10]. In our experiments, we set γ as 0.1 for the optimal value.

[1]The source code is available at https://github.com/pineapple0422/DLSL

TABLE II
THE mAP SCORES ON SUBCATEGORY-SPECIFIC TASK

Method	Img-to-Txt	Txt-to-Img	Avg
MHTN [29]	0.116	0.124	0.120
ACMR [30]	0.162	0.075	0.119
JRL [31]	0.160	0.190	0.175
GSPH [32]	0.140	0.179	0.160
CMDN [33]	0.099	0.123	0.111
SCA [34]	0.050	0.050	0.050
GXN [35]	0.023	0.035	0.029
FGCrossNet [3]	0.210	0.255	0.233
Our DLSL	**0.318**	**0.319**	**0.319**

TABLE III
THE mAP SCORES ON INSTANCE-SPECIFIC TASK

Method	Img-to-Txt	Txt-to-Img	Avg
FGCrossNet [3]	0.328	0.346	0.337
Our DLSL	0.341	0.349	0.345

III. EXPERIMENTS

A. Experimental Settings

Data Sources. Table I shows the statistics of data sources:

- **Data source 1 (for images): CUB_img** [36] contains 11 788 images of 200 subcategories, 5994 for training and 5794 for testing.
- **Data source 2 (for texts):** $PKUFG - XMedia_{txt}$ [3] contains 8000 texts of 200 subcategories, 4000 for training and 4000 for testing, which describe the subcategory information, such as habitat, eating habit, and background. Text paragraphs are extracted from several encyclopedia websites, such as Wikipedia.
- **Data source 3 (for texts): CUB_txt** [37] expands the **CUB_img** [36] by collecting fine-grained visual descriptions. It follows the division settings of **CUB_img** [36], containing 11 788 texts of 200 subcategories, 5994 for training and 5794 for testing. In each text, 10 single sentence visual descriptions are collected, which depict the appearance information of each instance, such as color and shape. Text sentences are collected through the Amazon Mechanical Turk platform.

Task Description. We categorize the fine-grained image-text retrieval into two tasks according to the granularity described by

TABLE IV
ABLATION STUDIES. ITA REFERS TO IMAGE-TEXT ALIGNMENT, AND DP REFERS TO DISCRIMINATIVE PROPERTY

Method	Subcategory-specific			Instance-specific		
	Img-to-Txt	Txt-to-Img	Avg	Img-to-Txt	Txt-to-Img	Avg
ResNet50 + BOW	0.210	0.255	0.233	0.328	0.346	0.337
ResNet50 + BOW / ITA	0.296	0.281	0.289	0.335	0.346	0.341
DLSL (ResNet50 + BOW / ITA+DP)	**0.318**	**0.319**	**0.319**	**0.341**	**0.349**	**0.345**

texts, that are **subcategory-specific** and **instance-specific**. In the **subcategory-specific** task, **Data source 1 + Data source 2** are used, where the text paragraphs focus on the subcategory information, such as habitat and eating habit. While in the **instance-specific** task, **Data source 1 + Data source 3** are employed, where the text sentences describe the appearance information of each instance, such as color and shape. Compared with the subcategory-specific task, the instance-specific task contains more variations and details. Since the description is made according to each image, the instance-specific task is more suitable for fine-grained cross domain retrieval.

Evaluation Metrics. We report the bi-directional performance: (1)img-to-txt (image query versus text gallery); (2)txt-to-img (text query versus image gallery). We adopt the mean Average Precision (mAP) as the evaluation metrics.

Implementation Details. All experiments are conducted on a 64-bit Ubuntu 16.04 with 2 Intel 2.40 GHz CPUs, 256 GB memory, and 6 NVIDIA Tesla GPUs. For fair comparison, we take the same features as the inputs. For images, we take 1024-dimensional CNN feature as input for all methods, which is extracted from the FC layer of ResNet50 [7]. For texts, we take the 1000-dimensional BOW [21] feature as the input.

The best values for α and β are chosen by five-fold cross-validation, and the scope is set in [0.001, 0.01, 0.1, 1, 10], as shown in Fig. 2. In experiments, we set α as 0.1, β as 0.1.

B. Experiment on Subcategory-Specific Task

Table II compares DLSL with established methods, including coarse-grained [29]–[35] and fine-grained [3] methods.

Comparison With Coarse-grained Methods. Among all the compared methods, the fine-grained methods generally outperform the coarse-grained methods [29]–[35]. The reason is that coarse-grained methods utilize global representations to express whole image and text. Such representations work well on coarse-grained scenario where objects from different classes are easy to be distinguished, but the performance will drop on fine-grained task since fine-grained task requires to distinguish different subcategories, whose distinctions are subtler.

Comparison With Fine-grained Methods. DLSL performs better than [3] in all the experiments. This benefits from the procedure of alignment, and it is contributed by the discriminative property of the dictionary learning technique. Since images and texts are associated through a common latent feature space, the corresponding sample variation is not very large. Thus the dictionary learned on the image samples can have a good reconstruction for the same subcategory of text samples. Meanwhile, inspired by the discriminative property, the representations in the latent space should be distinguishable enough. In other words, images or texts from different subcategories should be much more different in the latent space, thus the retrieval performance in the latent space can be improved further.

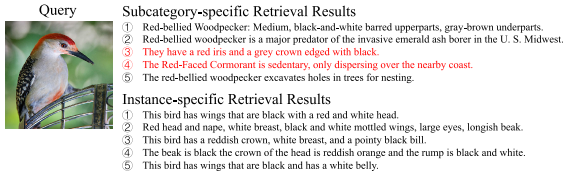

Fig. 3. An example on subcategory-specific and instance-specific image-to-text task. Red fonts are the wrong results.

C. Experiment on Instance-Specific Task

The experimental results on instance-specific task are shown in Table III. We can observe that the performance of DLSL outperforms FGCrossNet [3]. The reason is that great variation exists among each instances. The discriminative property of DLSL captures the variation to discriminate different fine-grained classes, so the alignment is more accurate. While FGCrossNet [3] just utilizes a uniform deep model for alignment, and ignores to make the alignment discriminative enough to adapt to the fine-grained task.

D. Ablation Study

Table IV shows the results of ablation studies. We use ResNet50 and BOW as the backbone. Note that **DP** refers to adding the explicit formulation for boosting the discriminative property. We observe that **DP** is important because it is designed to make the latent space more discriminative for the fine-grained task. Equipped with DP, our method improves the overall performance over the strongest competitor by an obvious margin (Avg: 0.289 $\rightarrow$ 0.319, 0.341 $\rightarrow$ 0.345).

E. Qualitative Results

To better understand the two specific fine-grained tasks, we visualize an example in Fig. 3. To be fair, we fix the query image and retrieval direction. We can see that the instance-specific task performs better than the subcategory-specific task. The reason is that large variation exists among the instances though they belong to the same subcategory. Since the text for the instance-specific task is generated based on specific image instances, the matching will be more accurate. While the description of text information in subcategory-specific task is fixed, they can not model the variations among different instances within the same subcategory.

IV. CONCLUSION

In this letter, we propose a novel DLSL for fine-grained image-text retrieval. DLSL first learns image and text features, and then performs couple dictionary learning to align the heterogeneous data in a uniform latent space, where the discriminative property is explicitly formulated for separating the subtle differences among fine-grained classes. We verify DLSL on two specific fine-grained tasks: subcategory-specific and instance-specific retrieval. Extensive experiments demonstrate the effectiveness of DLSL.

REFERENCES

[1] Y. Peng, X. Huang, and Y. Zhao, "An overview of cross-media retrieval: Concepts, methodologies, benchmarks, and challenges," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 9, pp. 2372–2385, Sep. 2018.

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

[3] X. He, Y. Peng, and L. Xi-e, "A new benchmark and approach for fine-grained cross-media retrieval," in *Proc. ACM Int. Conf. Multimedia*, 2019, pp. 1740–1748.

[4] A. Krizhevsky, I. Sutskever, and G. Hinton, "Imagenet classification with deep convolutional neural networks," *Adv. Neural Inf. Process. Syst.*, vol. 25, no. 2, pp. 1097–1105, 2012.

[5] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1–9.

[6] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Representations*, 2015, pp. 1–14.

[7] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[8] C. Deng, X. Xu, H. Wang, M. Yang, and D. Tao, "Progressive cross-modal semantic network for zero-shot sketch-based image retrieval," *IEEE Trans. Image Process.*, vol. 29, pp. 8892–8902, 2020, doi: 10.1109/TIP.2020.3020383.

[9] J. Zhang and Y. Peng, "SSDH: Semi-supervised deep hashing for large scale image retrieval," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 1, pp. 212–225, Jan. 2019.

[10] R. Wang, S. Qiao, S. Shan, and X. Chen, "Deep position-aware hashing for semantic continuous image retrieval," *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2020, pp. 2482–2491.

[11] J. Zhang and Y. Peng, "Query-adaptive image retrieval by deep-weighted hashing," *IEEE Trans. Multimedia*, vol. 20, no. 9, pp. 2400–2414, Sep. 2018.

[12] X. Tang, C. Deng, and X. Gao, "Discriminative latent feature space learning for cross-modal retrieval," in *Proc. ACM Int. Conf. Multimedia Retrieval*, 2015, pp. 507–510.

[13] C. Deng, X. Tang, J. Yan, W. Liu, and X. Gao, "Discriminative dictionary learning with common label alignment for cross-modal retrieval," *IEEE Trans. Multimedia*, vol. 18, no. 2, pp. 208–218, Feb. 2016.

[14] F. Shang, H. Zhang, L. Zhu, and J. Sun, "Adversarial cross-modal retrieval based on dictionary learning," *Neurocomputing*, vol. 355, pp. 93–104, 2019.

[15] J. Wu, Z. Lin, and H. Zha, "Joint dictionary learning and semantic constrained latent subspace projection for cross-modal retrieval," in *Proc. ACM Int. Conf. Inf. Knowl. Manage.*, 2018, pp. 1663–1666.

[16] H. Liu, F. Wang, X. Zhang, and F. Sun, "Weakly-paired deep dictionary learning for cross-modal retrieval," *Pattern Recognit. Lett.*, vol. 130, pp. 199–206, 2020.

[17] X.-S. Xu, "Dictionary learning based hashing for cross-modal retrieval," in *Proc. ACM Int. Conf. Multimedia*, 2016, pp. 177–181.

[18] X. Shen, F. Shen, Q. Sun, Y. H. Yuan, and H. Shen, "Robust cross-view hashing for multimedia retrieval," *IEEE Signal Process. Lett.*, vol. 23, no. 6, pp. 893–897, Jun. 2016.

[19] C. Zheng, L. Zhu, S. Zhang, and H. Zhang, "Efficient parameter-free adaptive multi-modal hashing," *IEEE Signal Process. Lett.*, vol. 27, no. 99, pp. 1270–1274, 2020, doi: 10.1109/LSP.2020.3008335.

[20] D. Xie, C. Deng, C. Li, X. Liu, and D. Tao, "Multi-task consistency-preserving adversarial hashing for cross-modal retrieval," *IEEE Trans. Image Process.*, vol. 29, pp. 3626–3637, 2020, doi: 10.1109/TIP.2020.2963957.

[21] L. Fei-Fei, R. Fergus, and P. Perona, "Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshop*, 2004, pp. 178–178.

[22] X. Zhang, S. Gui, Z. Zhu, Y. Zhao, and J. Liu, "Hierarchical prototype learning for zero-shot recognition," *IEEE Trans. Multimedia*, vol. 22, no. 7, pp. 1692–1703, Jul. 2020.

[23] E. Yang, C. Deng, C. Li, W. Liu, J. Li, and D. Tao, "Shared predictive cross-modal deep quantization," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 11, pp. 5292–5303, Nov. 2018.

[24] X. Zhang, Z. Zhu, Y. Zhao, and Y. Zhao, "Prolfa: Representative prototype selection for local feature aggregation," *Neurocomputing*, vol. 381, pp. 336–347, 2020.

[25] M. Zheng, Y. Jia, and H. Jiang, "Fine-grained image-text retrieval via complementary feature learning," in *Proc. Int. Conf. Multimedia Model.*, 2021, pp. 592–604.

[26] X. Zhang, Z. Zhu, Y. Zhao, D. Chang, and J. Liu, "Seeing all from a few: ℓ_1-norm-induced discriminative prototype selection," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 7, pp. 1954–1966, Jul. 2019.

[27] S. Hadfield, Z. Wang, B. O'Gorman, E. Rieffel, D. Venturelli, and R. Biswas, "From the quantum approximate optimization algorithm to a quantum alternating operator ansatz," *Algorithms*, vol. 12, no. 2, 2019, Art. no. 34.

[28] A. F. Martins, M. A. Figueiredo, P. M. Aguiar, N. A. Smith, and E. P. Xing, "An augmented lagrangian approach to constrained map inference," in *Proc. Int. Conf. Mach. Learn.*, 2011, pp. 2–10.

[29] 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.

[30] B. Wang, Y. Yang, X. Xu, A. Hanjalic, and H. Shen, "Adversarial cross-modal retrieval," in *Proc. ACM Int. Conf. Multimedia*, 2017, pp. 154–162.

[31] 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.

[32] D. Mandal, K. N. Chaudhury, and S. Biswas, "Generalized semantic preserving hashing for n-label cross-modal retrieval," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2633–2641.

[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] K.-H. Lee, X. Chen, G. Hua, H. Hu, and X. He, "Stacked cross attention for image-text matching," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 201–216.

[35] J. Gu, J. Cai, S. R. Joty, L. Niu, and G. Wang, "Look, imagine and match: Improving textual-visual cross-modal retrieval with generative models," *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7181–7189.

[36] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The Caltech-UCSD Birds-200-2011 dataset," Tech. Rep., 2011.

[37] S. Reed, Z. Akata, H. Lee, and B. Schiele, "Learning deep representations of fine-grained visual descriptions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 49–58.