IEEE
Computational
Intelligence
Society

Attribute-Guided Network for Cross-Modal Zero-Shot Hashing

Journal:	*IEEE Transactions on Neural Networks and Learning Systems*
Manuscript ID	TNNLS-2018-P-9024.R1
Manuscript Type:	Paper
Date Submitted by the Author:	16-Jul-2018
Complete List of Authors:	Ji, Zhong; Tianjin University, School of Electrical and Information Engineering Sun, Yuxin; Tianjin University, School of Electrical and Information Engineering Yu, Yunlong Pang, Yanwei; Tianjin University, School of Electronic Information Engineering Han, Jungong; Lancaster University, School of Computing & Communications
Keywords:	cross-modal hashing, zero-shot learning, attribute, image retrieval, zero-shot hashing

TNNLS-2018-P-9024. R1

Attribute-Guided Network for Cross-Modal Zero-Shot Hashing

Response to the Reviews

The authors thank the Associate Editor and all reviewers for the constructive comments, and have very carefully followed and cleaned up every raised issue. We provide below a detailed account on the changes that we have made in response to comments that the reviewers have provided. The corresponding changes in the R1 version are marked in *blue* color. Apart from addressing these concerns, we have also made some minor revisions. All these changed parts are labelled in *red* color.

To Reviewer # 1

1) In the last paragraph in Page 1, "It should be noted that Image-Based Text Retrieval (IBTR), …. , which is not the focus of our work", it is still not very clear to me. Please explain it more clearly.

Response: Thanks for your comment. The setting of IBTR in cross-modal retrieval is that the query set is constructed with images and the retrieval set is established with textual representations. However, under the current zero-shot cross-modal hashing setting, each category has only one textual representation. If we perform IBTR, the corresponding results in hunting scope is only one. Therefore, IBTR in this situation is actually degenerates into a ZSL, which is not the focus of our work.

2) Table II is suggested to present more detailed explanation for why it is set like that. For example, why using sigmoid function?

Response: Thanks for your constructive comments. It is based on the following considerations. Firstly, one of the properties of sigmoid function is that it is easy to reach saturation. Since the attribute annotations we used are binary labels, we use sigmoid function in the final layer in V2A and T2A net to ensure the prediction to fit the ground truth. Secondly, the combination of sigmoid function and cross-entropy cost function is widely used in deep neural network learning. Finally, the reason why we design AgNet in the architecture of fully connected network is that we have utilized high semantic features (GoogleNet and WordVec) as visual and textual features, the plain neural network is enough to generate effective hash codes.

3) Parameter scale and convergence speed of the model are suggested to be presented.

Response: Thanks for your constructive suggestion. We have added the description of parameter scale and the convergence curve of AgNet, which is as follow:

"The network architecture details of AgNet are shown in TABLE II. For each connection for different neurons, there are two parameters. Totally, AgNet consists of 1.3 million parameters."

"The convergence curve of AgNet is shown in Fig.3, AgNet reaches a stable objective function value at the 30-th epoch, which indicates the efficiency of AgNet."

TABLE II

The network architecture details of the proposed AgNet. "Full"denotes the fully-connected layer, "Relu" and "Sigmoid" denote activation functions.

Sub-network	Layer	Configuration
V2A Net	Full1+Relu Full2+Relu Full3+Sigmoid	1024 512 number of attributes d
T2A Net	Full1+Relu Full2+Sigmoid	1000 number of attributes d
A2H Net	Full1+Relu Full2+Sigmoid Full3	128 128 hash codes length c

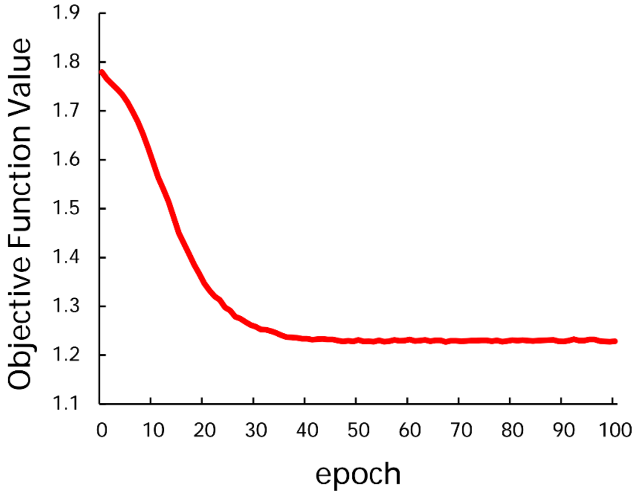


Fig. 3. Convergence curve of AgNet on AwA dataset.

4) Line 57-59, Page 4, fv() and ft() are not consistent with g().

Response: Thanks for pointing it out. We have unified the expression of these three items with $f_v(x_i;\theta_v)$, $f_t(z_i;\theta_t)$ and $g_v(\hat{a}_i;\theta_h)$.

5) The authors should discuss more comments about future work.

Response: Thanks for your constructive suggestion. We have discussed more comments about future work in the last section, which is as follows:

"In the future, as the acquisition of attribute annotation requires prior knowledge, we plan to exploit some other semantic information to formulate the common space, e.g., click-through data. We will also apply Hashlayer described in [40] to control the quantization error. In addition, we will exploit generative methods, e.g., GAN and VAE, to establish more robust embedding in zero-shot hashing."

To Reviewer # 2

1) In Section 3, the authors proposed a loss L_CS. Here, it is mentioned that "by minimizing this objective function, the Hamming distances for those instances within the same category but with different modalities are reduced." However, here Q and P have not been binarized yet. How can we conclude that minimizing L_CS leads to better alignment of different modalities in Hamming distance measure?

Response: Thanks for your comment. It is true that P and Q have not been binarized in the loss of L_cs. However, the purpose of a loss function is to reflect different relation intensities, which can be done by either real-valued variables or binary-valued variables. Actually, minimizing L_CS plays an indirect way to align different modalities in Hamming distance measure. Since P and Q are closely related to the final visual and textual hash codes, the operation on them can have an impact to the final result. Via minimizing the L_cs Loss, the products of P and Q reach to maximization when they are from the same categories, and reach to minimum when they are from different categories. In this way, it achieves a similar result comparing with that directly operating on the codes.

2) Again, this L_CS loss is the same as the first term of formula (1) in [Deep Cross-Modal Hashing]. Here, the authors should give reference to this paper. [Deep Cross-Modal Hashing] also denotes that the first term in (1) can preserve the cross-modal similarity in S with the image feature representation F and text feature representation G.

Response: Thanks for your suggestion. We have discussed the corresponding reference in Section III. C, which is as follow:

"Inspired by DCMH [21], we utilize the category similarity loss to ensure the different predicted attribute vectors of different modalities in the same category can generate similar hash codes, while those in different categories have distinct differences."

3) SUN dataset does not contain binary attributes, which contradicts what the authors said in Section 3.A. Actually, the proposed method does not require binary attributes. Thus, it is recommended to use real-valued attributes.

Response: Thanks for your constructive comment. SUN dataset has image-level attribute annotations and its attribute is real-value. We utilize binary class-level attribute annotations in AgNet with the following consideration:

a) We view the attribute prediction as the classification problem, where labels used in deep neural network are usually in the form of one-hot code.

b) Binary label increases the sparsity of annotation.

c) The calculation of attribute similarity involves class-level category similarity, which makes it necessary to transfer the image-level annotation into the class-level.

To get the class-level binary attribute annotations, we first calculate the average value for each category and then ceil the mean values to get the binary attributes.

4) In Section 3.C, when describing L_as, the S^(att) logic flow is somewhat vague. In this section, it is mentioned said "by minimizing this term, those instances closed in the attribute space and from different categories will be uncoupled in the Hamming space". This sentence seems to be vague as all the elements in this term are real-valued, which has nothing to do with hamming space. A more specific explanation is required to clarify the statement.

Response: Thanks for your comment. The reply to this comment is similar to that in comment 1. Although All elements in L_as are real-valued. We employ the inner products of these real-valued elements to capture relations among them, which has similar role with the binary-valued ones, but more accurate.

5) In Section4. C, SUN dataset was not included in single modality ZSH methods. However, SUN is one representative dataset whose category number is large while samples number is small. The authors should add experiment about SUN or clarify why not use SUN.

Response: Thanks for your comment. SUN is a representative dataset for zero-shot learning. However, SUN is not suitable for single modality ZSH setting. In the standard zero-shot split of SUN, the unseen class has 10 categories and each class has 20 instances. For the popular single modality ZSH methods. The query sets have 1,000 images from unseen class, and the retrieval set have more than 10,000 instances. The probe scale of SUN is too small to evaluate the effectiveness of the model. Therefore, SUN dataset is not suitable for single modality ZSH.

To Reviewer # 3

1) Two-stage v.s. joint optimization: AgNet is trained in two steps: firstly train V2A and T2A and then train A2T. I am wondering why not train the three subnetworks simultaneously (by controlling the gradient flow)? An experimental comparison should be performed between two-stage optimization and joint optimization.

Response: Thanks for your comment. Joint optimization is a good idea. However, in the proposed AgNet, the visual features and the text features are embedded into the common attribute space with V2A and T2A networks separately, and share the same A2H network. Thus, the gradient of the final objective function is hard to flow into V2A and T2A simultaneously. To this end, the whole network is hard to be optimized end-to-end. We will try it in our future work.

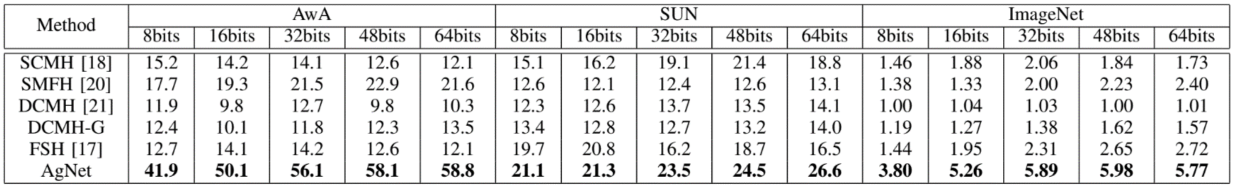
2) Among the comparison methods in this table, DCMH is the only deep method, whose backbone net of image feature extraction is variant of AlexNet. However, image feature extraction net of AgNet, as well as those

image features used in those shallow methods, are based on GoogleNet, which may lead to unfair comparison. It is reasonable to expect that the performance of DCMH can be improved while also using GoogleNet as image extraction network. Please include this variant into comparison.

Response: Thanks for your constructive suggestion. For fair comparison, we perform an additional experiment on DCMH by replacing the visual modality with GoogleNet, which is denoted as DCMH-G. The experimental results are in Section IV.B

TABLE III

Results on three benchmark datasets in Mean Average Precision (%) on CMZSH task.



Method	AwA					SUN					ImageNet				
	8bits	16bits	32bits	48bits	64bits	8bits	16bits	32bits	48bits	64bits	8bits	16bits	32bits	48bits	64bits
SCMH [18]	15.2	14.2	14.1	12.6	12.1	15.1	16.2	19.1	21.4	18.8	1.46	1.88	2.06	1.84	1.73
SMFH [20]	17.7	19.3	21.5	22.9	21.6	12.6	12.1	12.4	12.6	13.1	1.38	1.33	2.00	2.23	2.40
DCMH [21]	11.9	9.8	12.7	9.8	10.3	12.3	12.6	13.7	13.5	14.1	1.00	1.04	1.03	1.00	1.01
DCMH-G	12.4	10.1	11.8	12.3	13.5	13.4	12.8	12.7	13.2	14.0	1.19	1.27	1.38	1.62	1.57
FSH [17]	12.7	14.1	14.2	12.6	12.1	19.7	20.8	16.2	18.7	16.5	1.44	1.95	2.31	2.65	2.72
AgNet	**41.9**	**50.1**	**56.1**	**58.1**	**58.8**	**21.1**	**21.3**	**23.5**	**24.5**	**26.6**	**3.80**	**5.26**	**5.89**	**5.98**	**5.77**

We utilize GoogleNet as backbone net in DCMH-G. As illustrated in Table III, we observe that DCMH-G outperforms DCMH on three datasets in most cases. However, AgNet clearly outperforms both DCMH and DCMH-G on three benchmark datasets, consistently.

3) Zero-shot cross-modal retrieval or just classification? If I understand correctly, text fed into T2A is just name of category (single word for each image). This is ad-hoc because text data of most cross-modal benchmarks is caption (MSCOCO, Flicker-8K, etc) or article (wikipedia). The cross-modal retrieval task performed here actually is classifying images into category (via nearest neighbor between the hash codes of image and category name), which largely reduces the difficulties in general cross-modal retrieval (e.g. complex semantics and noises in text descriptions) and makes such attempt becomes trivial. Personally speaking, considering the label and class-attribute predicate are given, attribute names of each image might be used as the text modality, which have richer information than just name of category.

Response: Thanks for your constructive comments. Actually, retrieval and classification can be viewed as inverse procedures. Both retrieval and classification are trying to establish the relationship between visual samples and class prototypes. The difference between classification and retrieval is that classification aims at categorizing samples into class prototypes, and retrieval attempt to find the Top-N instances which are relate to the samples.

In the setting of traditional cross-modal retrieval, each class has plenty of textual representations, which indeed increase the difficulty in retrieval. Nevertheless, the representative zero-shot benchmarks are short of enough textual representation as cross-modal datasets. On the other hand, the textual representations from cross-modal benchmarks are hard to split under the zero-shot setting, where the categories of seen data and unseen data are non-intersect. In addition, using the names of categories as a query set is the popular setting in current zero-shot retrieval work [1][2][3].

And using the attribute names as the textual representations is not suitable for the current zero-shot cross-modal setting, as the attributes are shared with seen classes and unseen classes, the names of attribute can't be effectively split under the zero-shot standard.

4) In part Ⅵ.C, AgNet is compared with previous zero-shot hashing schemes and hashing methods for non zero-shot setting. However, it is quite natural to consider an abbreviated version of AgNet: which is only consisted of V2A and A2H. I am wondering if incorporating T2A into AgNet is beneficial to SMZSH? i.e., Does mitigating modality gap between image and text contributes to the zero-shot binary codes learning of images? Considering the significance of CMZSH task is quite doubtful (discussed in (3)), more experiments should be done in this part to reveal some "mutual promotion" for enhancing the significance. Otherwise this work is just a brutal-force combination of cross-modal hashing and zero-shot hashing, which will largely limit its novelty.

Response: Thanks for pointing this issue out. In order to verify the contributions of each modality (e.g., visual and text), we conduct a list of ablation experiments on large-scale ImageNet dataset, the experimental details are in Part IV.C.3).

TABLE IV

Performances (mAP / Precision) of AgNet and AgNet-vis on ImageNet dataset on CMZSH task

method	8bits	16bits	32bits	48bits
AgNet-vis	4.33/2.41	5.66/5.11	6.92/10.7	8.22/15.71
AgNet	4.40/2.80	6.09/7.31	6.94/13.3	8.46/17.51

We implement a visual version (denoted as AgNet-vis) of AgNet on ImageNet dataset. The main different between AgNet and AgNet-vis is that AgNet-vis removes the T2A-Net. In addition, to maintain the category similarity, AgNet-vis replaces $\Theta_{ij} = \frac{1}{2} P_{*i}{}^{T} Q_{*j}$ with $\phi_{ij} = \frac{1}{2} P_{*i}^{T} P_{*j}$ in category similarity loss function. The experimental results (mAP / Precision) on ImageNet is shown in Table IV. It can be observed that AgNet outperforms AgNet-vis in all code lengths consistently, which demonstrates that mitigating modality gap between image and text contributes to the zero-shot binary codes learning of images

5) Page.7 line 6-9: the authors claim that SitNet uses random split while AgeNet uses standard split. But how to fairly compare these two methods is still unsolved.

Response: Thanks for pointing it out. For fair comparison, we randomly split the seen and unseen domain following the SitNet and individually compare AgNet and SitNet in Fig. 5. The result is shown as follow:

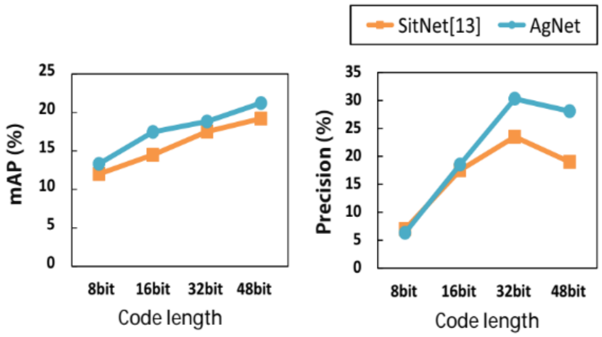


Fig. 5. Performances (mAP and Precision) of SitNet and AgNet for singlemodal zero-shot hashing task on AwA dataset.

As illustrated in Fig.5. AgNet outperforms SitNet in most situations. For instance, the mAP of AgNet gains 18.8% on 32 bits, which has an improvement against SitNet by 7.4% in the same code length.

6) As for the experimental setting, I would like to throw doubt on the way of constructing query and retrieval sets described in line 15~20 in Page.6, which is derived from [4]. In zero-shot classification, classfying samples from U and classifying samples from S∪U are defined as ZSL and Genaralized ZSL (GZSL), respectively [6]. The following two settings might be experimented separately: a) query and retrieval sets are both from unseen images; b) are mixture of seen and unseen images.

Response: Thanks for your constructive comment. Traditional ZSL (TZSL) and GZSL are two primary tasks in Zero-Shot Learning, where GZSL is more challenging than TZSL. This is because that GZSL enlarges the test set with the whole dataset and increases the obstruction of seen data. Besides, SitNet[5] constructs its experimental setting following GZSL criterion. Therefore, we follow the setting of SitNet in AgNet and believe GZSL is enough to evaluate the performance of model.

7) Page.2 line 19: the generated hash codes should not be searched by nearest neighbor. Instead, approximate nearest neighbor (ANN) is used.

Response: Thanks for pointing it out. We have replaced nearest neighbor with approximate nearest neighbor, which is as follow:

"To implement effective approximate nearest neighbor (ANN) search, the generated binary hash codes are necessary to inherit the semantic similarity relationship of high dimensional real-value features."

8) Hashlayer described in [7] might be a good plug-in replacement to the way of controlling quantization error used in this paper.

Response: Thanks for your constructive suggestion. Hashlayer in [7] is a remarkable work to figure out the vanishing gradient problem of sign function. During the optimization stage in AgNet, we use the inner product between the outputs of network to maintain the category and attribute similarity in Hamming Space. Besides, the outputs of sign function in our algorithm act as a regularization item without suffering from vanishing gradient problem. We believe the Hashlayer in [7] is enlightening and beneficial for our future work, which has been discussed in the last section.

References

[1] J. Lu, J. Li, Z. Yan, and C. Zhang, “Zero-Shot Learning by Generating pseudo feature representations,” arXiv:1703.06389, 2017.

[2] X. Xu, F. Shen, Y. Yang, et al, “Matrix tri-factorization with manifold regularizations for zero-shot learning,” in Proc. IEEE Conf. on Comput. Vis. Pattern Recognit., Honolulu, Hawaii, USA, July, 2017.

[3] M. Bucher, S. Herbin, and F. Jurie, “Improving Semantic Embedding Consistency by Metric Learning for Zero-Shot Classiffication,” in Eur. Conf. Comput. Visi., Amsterdam, Netherlands, Oct. 2016, pp. 730-746.

[4] Y. Yang, Y. Luo, L. Chen, F. Shen, J. Shao, and H. Shen, “Zero-shot hashing via transferring supervised knowledge,” in Proc. ACM Conf. on Multimedia, Amsterdam, Netherlands, Oct. 2016, pp. 1286-1295, 2016.

[5] Y. Guo, G. Ding, J. Han, and Y. Guo, “SitNet: discrete similarity transfer network for zero-shot hashing,” in Int. Joint Conf. on Art. Intell., Melbourne, Australia, Aug. 2017, pp. 1767-1773.

[6] W-L. Chao, S. Changpinyo, B. Gong, F. Sha, "An Empirical Study and Analysis of Generalized Zero-Shot Learning for Object Recognition in the Wild," in Eur. Conf. Comput. Visi., Amsterdam, Netherlands, Oct. 2016.

[7] Z. Cao, M. Long, J. Wang, P. S. Yu, "HashNet: Deep Learning to Hash by Continuation", in Int. Conf. Comput. Visi. 2017.

Attribute-Guided Network for Cross-Modal Zero-Shot Hashing

Zhong Ji, *Member, IEEE,* Yuxin Sun, Yunlong Yu, Yanwei Pang, *Senior Member, IEEE,* Jungong Han

***Abstract*—Zero-Shot Hashing aims at learning a hashing model that is trained only by instances from seen categories but can generate well to those of unseen categories. Typically, it is achieved by utilizing a semantic embedding space to transfer knowledge from seen domain to unseen domain. Existing efforts mainly focus on single-modal retrieval task, especially Image-Based Image Retrieval (IBIR). However, as a highlighted research topic in the field of hashing, cross-modal retrieval is more common in real world applications. To address the Cross-Modal Zero-Shot Hashing (CMZSH) retrieval task, we propose a novel Attribute-Guided Network (AgNet), which can perform not only IBIR, but also Text-Based Image Retrieval (TBIR). In particular, AgNet aligns different modal data into a semantically rich attribute space, which bridges the gap caused by modality heterogeneity and zero-shot setting. We also design an effective strategy that exploits the attribute to guide the generation of hash codes for image and text within the same network. Extensive experimental results on three benchmark datasets (AwA, SUN, and ImageNet) demonstrate the superiority of AgNet on both cross-modal and single-modal zero-shot image retrieval tasks.**

***Index Terms*—Zero-shot hashing, cross-modal hashing, zero-shot learning, attribute, image retrieval.**

I. Introduction

RECENTLY, hashing-based multimedia retrieval approaches have attracted a lot of attention, mainly owing to their fast retrieval speed and low storage cost [1], [2], [3]. Generally, these approaches fall into two categories: unsupervised hashing [1], [2], [4], [5] and supervised hashing [6], [7]. The former usually applies the statistics information, such as manifold structure [4] and the variance of feature [5], to generate the hash function with the intention to preserve the similarity space, while the latter explores the semantic supervision information, *e.g.*, class label, to capture the intrinsic property of data. Because more knowledge is utilized, supervised hashing approaches usually achieve better performance than those of unsupervised ones. However, one deficiency of supervised hashing approaches is that a large number of labeled instances are required for training the model, which is time-consuming and labor-intensive. In addition, it is very difficult to annotate sufficient training data for the new concepts in a timely manner, and also, impractical to retrain the hashing model whenever the retrieval system meets a new concept [12].

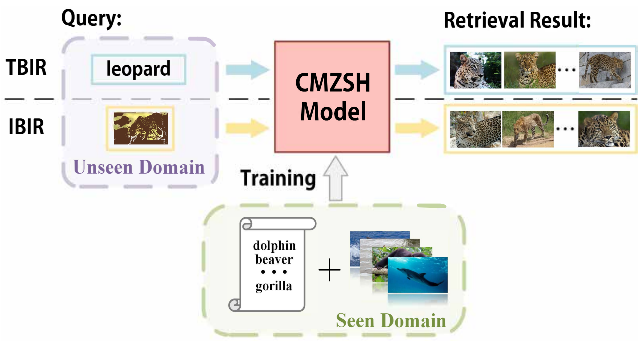


Fig. 1. An illustration of Cross-Modal Zero-Shot Hashing (CMZSH). Typically, a CMZSH model is trained by texts and images in seen domain. At testing stage, the CMZSH model mainly tackles two tasks in unseen domain, *i.e.,* Text-Based Image Retrieval (TBIR) and Image-Based Image Retrieval (IBIR). For TBIR, the query set are texts and the retrieval set are images. For IBIR, both the query and the retrieval sets are images.

To address this awkward situation, inspired by the success of Zero-Shot Learning (ZSL) [8], [9], [10], [11], Zero-Shot Hashing (ZSH) is developed recently [12], [13]. Its goal is to encode images of unseen categories with the hash funciton trained by only those of seen categories by incorporating the ideas of supervised hashing approaches and ZSL. Transferring Supervised Knowledge (TSK) [12] is the pioneering method in ZSH. The authors propose to employ the semantic vectors as a bridge to transfer available supervision information from seen categories to unseen categories. Further, Guo *et al*. [13] present a deep ZSH method, named Similarity Transfer Network (SitNet). Specifically, SitNet applies a multi-task architecture to leverage the supervision knowledge of seen categories and the semantic vectors simultaneously, and employs a straight-through estimator to avoid information loss caused by real-value relaxation. Although these methods have achieved impressive performance, there is still a serious limitation for them. That is, the existing ZSH approaches only focus on Image-Based Image Retrieval (IBIR) task, where both the query and the retrieval sets are images. In fact, Text-Based Image Retrieval (TBIR), *i.e.*, leveraging textual description to search images, is also very common in the real-life scenario.

The aforementioned limitation motivates us to consider investigating ZSH in a cross-modal retrieval setting, which we call Cross-Modal Zero-Shot Hashing (CMZSH). Specifically, CMZSH mainly deals with two different tasks, one is IBIR, and the other is TBIR. That is to say, CMZSH broadens the scope of conventional ZSH from single-modal application to cross modal application. An illustration is described in Fig. 1. It should be noted that Image-Based Text Retrieval (IBTR)

This work was supported by the National Natural Science Foundation of China (Grant No. 61771329 and No. 61632018), and the National Basic Research Program of China (Grant No. 2014CB340400).

Z. Ji (corresponding author), Y. Sun, Y. Yu (corresponding author), and Y. Pang are with the School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China (e-mail: {jizhong, sunyuxin, yuyunlong, pyw}@tju.edu.cn).

J. Han is with the School of Computing & Communications, Lancaster University, UK (e-mail: jungong.han@lancaster.ac.uk).

also belongs to the scope of Cross-Modal Hashing. However, since only one category name is corresponding to a class of images for most popular ZSH datasets, IBTR in this situation actually degenerates into a ZSL (also called zero-shot image classification) problem, which is not the focus of our work.

To achieve CMZSH, the following challenges should be addressed. 1) Modality heterogeneity. As query set and retrieval set are likely to be from different modalities, the generated hash codes are expected to have an additional property that preserves the semantic relationship between both modalities. 2) Category migration. It is an inherent problem of ZSL that the learning model should have the ability of handling the instances from unseen categories. Therefore, CMZSH needs to exploit the transferable knowledge that bridges the gap between seen categories and unseen categories. 3) Semantic similarity preservation. The hash function is actually a projection from high dimensional real-value features to low dimensional binary space. To implement effective approximate nearest neighbor (ANN) search, the generated binary hash codes are necessary to inherit the semantic similarity relationship of high dimensional real-value features.

In this paper, we address the above issues with the proposed Attribute-Guided Network (AgNet) framework. Specifically, to narrow the semantic gap brought by modality heterogeneity and category migration, we map both the visual features and the textual features into a common space, respectively. In this work, we utilize the class-level attribute space as the common space. In this way, the two different modalities are aligned into a high-level semantic space. Using the embeddings of different modalities in the attribute space as inputs, both visual and textual hash codes are obtained from a shared deep neural network. Besides, the relationships between different modalities are constructed via a category similarity matrix formulated with the pair-wise class label. Moreover, to preserve the relationship of different categories, attribute similarity is further introduced to restrict the distances of different categories in the same modality.

We summarize our highlights as below:

1) We address the cross-modal retrieval problem in ZSH, *i.e.*, Cross-Modal Zero-Shot Hashing (CMZSH), via a novel deep hashing neural network. It can perform not only IBIR, but also TBIR. To the best of our knowledge, it is the first work to study the cross-modal hashing retrieval in the zero-shot setting.
2) By exploiting the class-level attributes information, we propose an Attribute-Guided Network (AgNet) framework. It first maps two different modalities into a common attribute space, which acts as a hub to bridge unseen and seen categories, as well as visual and textual modalities. Then, an effective strategy is designed to generate two individual hash codes for image and text within the same network. Specifically, we exploit the attribute to guide the generation of hash codes by preserving the category similarity and attribute similarity.
3) The experimental results for both IBIR and TBIR tasks on three popular benchmark datasets demonstrate that the proposed AgNet achieves competitive performance.

II. Related Work

In this section, we will introduce some research progresses closely related to our work, including cross-modal hashing and zero-shot hashing. In fact, CMZSH can be viewed as a special case for them. CMZSH also falls into the domains of hashing-based retrieval and zero-shot learning. Due to the limited space, please refer to [14] and [15] for more elaborate surveys about them.

A. Cross-Modal Hashing

Cross-Modal Hashing (CMH) is a widely used retrieval technique [3], [16], [18], most of which tackle the problems of Text-Based Image Retrieval (TBIR) and Image-Based Text Retrieval (IBTR). This is usually implemented by generating two respective hash codes for each individual modality. In this way, different modalities can be computed in the same hashing space. A number of methods have been proposed, which can be generally divided into two categories: unsupervised methods and supervised methods. As one of the representative unsupervised cross-modal methods, Collective Matrix Factorization Hashing (CMFH) [16] generates cross-modal hash codes in a latent semantic space shared by both modalities via collective matrix factorization technique. To explore the heterogeneous correlation in different modalities, Liu *et al.* [17] propose a novel CMH scheme using fusion similarity from the multiple modalities.

On the other hand, supervised CMH methods usually perform better than unsupervised ones since they can fully utilize the intrinsic property in data. For example, Zhang *et al.* [18] merge semantic labels into hashing learning procedure and propose a Semantic Correlation Maximization Hashing (SCMH) method. Lin *et al.* [19] convert the semantic similarity of instances into a probability distribution and generate hash codes by minimizing the KL-divergence. Liu *et al.* [20] propose a graph-regularized Supervised Matrix Factorization Hashing (SMFH) framework with a collective non-negative matrix factorization technique. With the renaissance of the deep neural network, deep learning has proved its outperformance in this field. For instance, Jiang *et al.* [21] first propose an end-to-end deep neural network framework to address the CMH problem. However, they just utilize the inter-modal relationship but ignore intra-modal information. To address this problem, Yang *et al.*[22] use pairwise labels to exploit intra-modal similarity and propose a Pairwise Relationship Guided Deep Hashing (PRDH) method.

Our proposed CMZSH framework follows the idea of supervised CMH, which leverages semantic supervision information to generate different hash codes for each modality to ensure they are able to interact with each other. However, different from CMH, CMZSH has to tackle an additional zero-shot problem. That is, the supervision knowledge is limited to seen categories, which is the only information in learning reliable hash function for transforming modalities of unseen categories into binary codes. Therefore, CMZSH is more challenging.

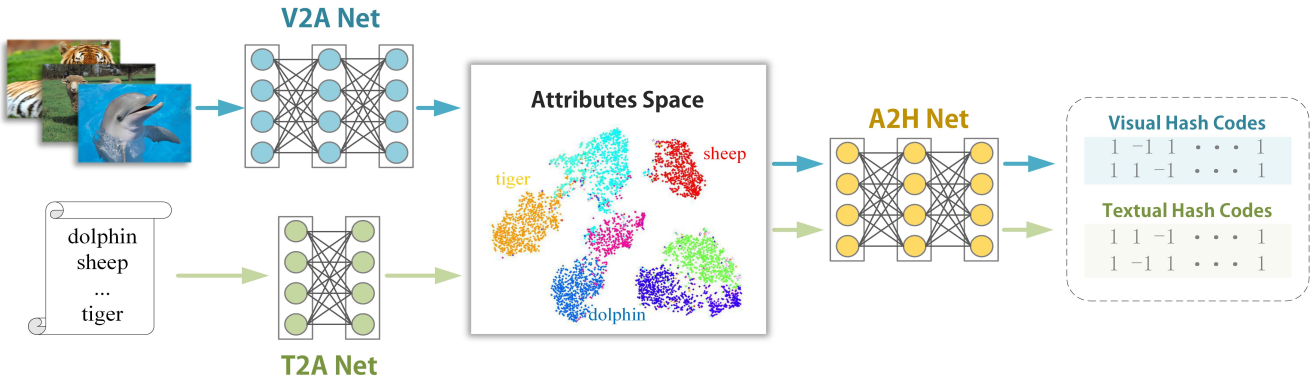


Fig. 2. Architecture overview of the proposed AgNet approach. It consists of two stages. First, V2A Net and T2A Net embed the inputs of image and text into a shared attribute space, respectively. Next, A2H Net encodes the visual and textual vectors in attribute space into visual hash codes and textual hash codes, respectively. The shared attribute space enables the knowledge transferability from seen categories to unseen categories. And the A2H Net makes the cross-modal retrieval feasible.

B. Zero-Shot Hashing

Zero-Shot Hashing (ZSH) is a marriage of zero-shot learning and hashing-based retrieval techniques. It is proposed to tackle the close-set limitation in hashing-based retrieval approaches, *i.e.*, the concepts of possible testing instances in either dataset or query set are within the training set [12], [13]. Therefore, ZSH explores only the information from seen categories to build hash functions to retrieve the images in unseen categories.

As an emerging research topic, the existing ZSH methods mainly focus on IBIR task. For example, in the pioneering work proposed by Yang *et al.* [12], the labels of each seen category are converted into semantic embedding representations via word2vec model [23], by which the supervision knowledge in seen categories can be transferred to unseen ones. Then, hash codes are generated by projecting the visual representation to the embedding space. Instead of using word vector as semantic representation in [12], Xu *et al.* [24] adopt semantically-rich attribute information as transferable knowledge. Further, Guo *et al.* [13] propose a multi-task framework to simultaneously exploit the supervision information from visual concepts and semantic representations. Specifically, they leverage the hash codes to capture the semantic similarity relationship in a transferable semantic embedding space and propose a center regularization loss to preserve both intra-concept similarity and inter-concept distance. In addition, under the transductive setting [25], [26], Lai *et al.* [27] propose a transductive zero-shot hashing method via coarse-to-fine similarity mining. In this way, a greedy binary classification network is first used to detect the most informative images from unseen category images. After that, the fine similarity mining module further finds the similarities among the informative images. However, since these ZSH approaches are designed for IBIR task, they have a natural deficiency that cannot encode the text into hash codes. Therefore, they are hardly applied for TBIR.

To achieve CMZSH, the idea of ZSH should be combined with that of CMH. This is exactly what this paper is going to tackle.

TABLE I
The main notations.

Notation	Description
N	number of instances
s	number of seen categories
u	number of unseen categories
d	number of attributes
c	hash codes length
l	dimensionality of visual space
k	dimensionality of textual space
$\mathbf{x} \in \mathbb{R}^l$	visual representation vector
$\mathbf{z} \in \mathbb{R}^k$	textual representation vector
$\mathbf{y} \in \mathbb{R}^{s+u}$	label vector
$\mathbf{a} \in \mathbb{R}^d$	ground-truth attribute vector
$\hat{\mathbf{a}}^{(v)} \in \mathbb{R}^d$	predicted attribute vector in visual modality
$\hat{\mathbf{a}}^{(t)} \in \mathbb{R}^d$	predicted attribute vector in textual modality
$\mathbf{S}^{(c)} \in \mathbb{R}^{n\times n}$	category similarity matrix
$\mathbf{S}^{(att)} \in \mathbb{R}^{n\times n}$	attribute similarity matrix
$\mathbf{P} \in \mathbb{R}^{c\times n}$	outputs matrix of A2H Net in visual modality
$\mathbf{Q} \in \mathbb{R}^{c\times n}$	outputs matrix of A2H Net in textual modality
$\mathbf{B} \in \mathbb{R}^{c\times n}$	hash codes matrix

III. The Proposed AgNet Algorithm

A. Problem Definition

In order to address the CMZSH problem, both the requirements of knowledge transferability from seen categories to unseen categories and cross-modal retrieval should be fulfilled. Attributes and word vectors are two most popular side information in ZSL [28], [29], [30], [31]. Specifically, attributes define a few properties of an object, such as color, shape, and the presence or absence of a certain body part. They are shared across both seen and unseen categories. Word vectors represent words as vectors based on distributed language representation techniques, and theoretically, they can encode arbitrary concepts into sematic vectors. Therefore, both attributes and word vectors can construct a semantic space to transfer the knowledge from seen categories to unseen categories, meaning either of them can be selected as the candidate semantic space in CMZSH. Further, different approaches can be designed to generate both visual and textual hash codes from either attributes or word vectors space, which enables the cross-modal retrieval feasible. In this paper, we only focus on the usage of attributes. That is to say, we exploit attributes as the

TABLE II
The network architecture details of the proposed AgNet. “Full”denotes the fully-connected layer, “Relu” and “Sigmoid” denote activation functions.

Sub-network	Layer	Configuration
V2A Net	Full1+Relu	1024
	Full2+Relu	512
	Full3+Sigmoid	number of attributes d
T2A Net	Full1+Relu	1000
	Full2+Sigmoid	number of attributes d
A2H Net	Full1+Relu	128
	Full2+Sigmoid	128
	Full3	hash codes length c

intermediary space, from which the hash codes are encoded.

Suppose we are given N training instances $\mathcal{D}_{tr} = \{d_i = (\mathbf{x}_i, \mathbf{z}_i, \mathbf{y}_i), i = 1, ..., N\}$ from s labeled seen categories $S = \{1, 2, ...s\}$, where $\mathbf{x}_i \in \mathbb{R}^l$ is the visual representation, $\mathbf{z}_i \in \mathbb{R}^k$ is the textual semantic representation of its corresponding category name and $\mathbf{y}_i \in \{0, 1\}^s$ is the label vector represented as one-hot encoding. Note that the different modalities mainly refer to image and text in this paper. Besides, each instance is also annotated with a binary attribute vector denoted as $\mathbf{a}_i \in \{0, 1\}^d$. Under the zero-shot setting, there also exist unseen categories $U = \{s + 1, ..., s + u\}$, which is disjoint with the labeled seen categories, *i.e.*, $S \cap U = \emptyset$.

B. Network Architecture

The overall framework of the proposed AgNet framework is illustrated in Fig. 2. It consists of three components: i) **V2A Net**. The output of penultimate layer (before the Softmax layer) of the fine-tuned GoogleNet [32] is first extracted as the visual features. After that, these CNN features are utilized as the input to a deep neural network with three fully-connected layers, which embeds the visual features to attribute vectors. ii) **T2A Net**. We use word2vector model [23] to represent the text input, which has been trained on the Wikipedia corpus. It is a 1000-dimensional vector for each category name. T2A Net is a two-layer neural network that is used to establish the word vectors to attributes projection. iii) **A2H Net**. Unlike the existing deep cross-modal hashing methods that generate hash codes from two independent networks (one for image, and the other for text), AgNet accomplishes the hash codes generation only with a single three-layer neural network. Specifically, it utilizes the predicted attribute vectors (or called attribute embedding vectors) as input, and outputs both visual and textual hash codes. Table II shows the configuration of AgNet. AgNet consists of 1.3 million parameters. It needs to be highlighted that the architecture of the neural network is not the focus of this work, what we want to prove is that attribute-guided framework is reasonable and beneficial for the performance of CMZSH.

C. Objective Function

We first design the objective functions for the V2A Net and the T2A Net, whose purpose is to transform the inputs of image and text to the attribute space. Their transformation functions are denoted as f_v and f_t, respectively. Let $\hat{\mathbf{a}}_i^{(v)} = f_v(\mathbf{x}_i; \theta_v) \in \mathbb{R}^d$ denote the predicted attribute vector of each visual representation $\mathbf{x}_i$ while $\hat{\mathbf{a}}_i^{(t)} = f_t(\mathbf{z}_i; \theta_t) \in \mathbb{R}^d$ denotes the predicted attribute vector of each textual representation $\mathbf{z}_i$. Given a training set of instances and their corresponding category attribute vectors, the V2A Net and the T2A Net are both trained with the cross-entropy objective function:

$$\mathcal{L}_{att} = -\frac{1}{N}\sum_{i=1}^{N} \mathbf{a}_i{}^T \log(\hat{\mathbf{a}}_i) + (\mathbf{1} - \mathbf{a}_i)^T \log(1 - \hat{\mathbf{a}}_i), \quad (1)$$

where $\mathbf{a}_i$ denotes the attribute vector, and $\hat{\mathbf{a}}_i$ is the predicted attribute vector $\hat{\mathbf{a}}_i^{(v)}$ for V2A Net or $\hat{\mathbf{a}}_i^{(t)}$ for T2A Net. This objective function ensures that the predicted attribute vectors approximate to the distribution of original attribute vectors.

Then, the key challenge is how to realize the purpose of A2H Net, *i.e.*, to generate two individual hash codes for image and text from the attribute space. We design three functions to achieve this purpose: i) category similarity loss; ii) attribute similarity loss; and iii) regularization loss.

Inspired by DCMH [21], we utilize the category similarity loss to ensure the different predicted attribute vectors of different modalities in the same category can generate similar hash codes, while those in different categories have distinct differences. Given the predicted attribute vectors$\hat{\mathbf{a}}^{(v)}$ and $\hat{\mathbf{a}}^{(t)}$ of image and text, respectively, denote $\mathbf{P}_{*i} = g\left(\hat{\mathbf{a}}_i^{(v)}; \theta_h\right) \in \mathbb{R}^c$ and $\mathbf{Q}_{*i} = g\left(\hat{\mathbf{a}}_i^{(t)}; \theta_h\right) \in \mathbb{R}^c$ as their outputs of A2H Net, where g is the transformation function for A2H Net and θ is the parameters for it. Moreover, use $\mathbf{\Theta}_{ij} = \frac{1}{2}\mathbf{P}_{*i}{}^T\mathbf{Q}_{*j}$ to represent the neighbor relationship between $\mathbf{P}_{*i}$ and $\mathbf{Q}_{*j}$ in the Hamming space . Denote $\mathbf{S}^{(c)} \in \mathbb{R}^{n \times n}$as the category similarity, where $\mathbf{S}_{ij}^{(c)} = 1$ when $\mathbf{y}_i = \mathbf{y}_j$ and $\mathbf{S}_{ij}^{(c)} = 0$ otherwise. By using the negative log likelihood of the inter-modal similarities, we formulate the category similarity loss as:

$$\mathcal{L}_{cs} = -\sum_{i,j=1}^{N} \left(\mathbf{S}_{ij}^{(c)}\mathbf{\Theta}_{ij} - \log\left(1 + e^{\mathbf{\Theta}_{ij}}\right)\right). \quad (2)$$

By minimizing this objective function, the Hamming distances for those instances within the same category but with different modalities are reduced, whereas the distances are getting larger for those with different categories. Therefore, the category similarity is preserved between different modalities.

In addition, an effective hash code should also be equipped with the discriminative ability in a single modality. Hence, attribute similarity matrix $\mathbf{S}^{(att)}$ is introduced to make the intra-modal hash codes more discriminable. Let $\mathbf{S}_{ij}^{(att)} = \cos(\mathbf{a}_i, \mathbf{a}_j) - \mathbf{S}_{ij}^{(c)}$, where $\cos(\mathbf{a}_i, \mathbf{a}_j)$ is the cosine distance between attribute vectors of $\mathbf{a}_i$ and $\mathbf{a}_j$. $\mathbf{S}^{(att)}$is a modified cosine similarity, which is used to measure the semantic similarities among different categories. Different from the binary label similarity $\mathbf{S}^{(c)}$, $\mathbf{S}^{(att)}$ utilizes a real-value to describe the similarities among different categories. It is used in the attribute similarity loss $\mathcal{L}_{as}$ as a guide for the generation of visual hash codes. Specifically, if two attribute vectors from different categories are similar, their corresponding visual instances should be given a higher penalty such that their hash codes have higher discriminative ability. If two instances are from the same category, we do not give them penalty, that is why $\mathbf{S}^{(c)}$ is subtracted. The attribute similarity loss is defined as follow:

$$\mathcal{L}_{as} = \sum_{i,j=1}^{N} \sigma\left(\phi_{i,j}\mathbf{S}_{i,j}^{(att)}\right), \tag{3}$$

where $\phi_{ij} = \frac{1}{2}\mathbf{P}_{*i}^{T}\mathbf{P}_{*j}$ represents the neighbor relationship of $\mathbf{P}_{*i}$ and $\mathbf{P}_{*j}$ in Hamming space, and $\sigma(\bullet)$ denotes the sigmoid function. The sigmoid function is applied to restrict the scope of this term. By minimizing this term, those instances closed in the attribute space and from different categories will be uncoupled in the Hamming space.

Meanwhile, we use $\|\mathbf{B}-\mathbf{P}\|_F^2$ to make $\mathbf{P}$ approximate to hash codes. And $\left\|\mathbf{P}^T\mathbf{1}\right\|_F^2$ ensures each bit of the hash codes is balanced. Then, the regularization loss is formulated as follow:

$$\mathcal{L}_{reg} = \|\mathbf{B}-\mathbf{P}\|_F^2 + \left\|\mathbf{P}^T\mathbf{1}\right\|_F^2, \tag{4}$$

where $\mathbf{B} = sign(\mathbf{P})$, $\mathbf{1}$ denotes a vector with all elements being 1.

Therefore, the overall objective function of the A2H Net is written as follow:

$$\begin{aligned}\mathcal{L}_{A2H} &= \mathcal{L}_{cs} + \lambda\mathcal{L}_{as} + \eta\mathcal{L}_{reg}\\ &= -\sum_{i,j=1}^{N}\left(\mathbf{S}_{ij}^{(c)}\mathbf{\Theta}_{ij} - \log\left(1+e^{\mathbf{\Theta}_{ij}}\right)\right)\\ &+ \sum_{i,j=1}^{N}\lambda\sigma\left(\phi_{i,j}\mathbf{S}_{i,j}^{(att)}\right) + \eta\left(\|\mathbf{B}-\mathbf{P}\|_F^2 + \left\|\mathbf{P}^T\mathbf{1}\right\|_F^2\right),\end{aligned} \tag{5}$$

where λ and η are trade-off parameters to control the weight of each item.

D. Optimization

Our AgNet is trained in two steps. Firstly, V2A Net and T2A Net are separately learned with cross entropy functions. Then, using the predicted attribute vectors from two modalities, we train A2H Net according to Eq. (5). Back Propagation algorithm is adopted to optimize AgNet. For Eq.(5), the gradient of $\frac{\mathcal{L}_{A2H}}{\partial\mathbf{P}_{*i}}$ is calculated with:

$$\begin{aligned}\frac{\mathcal{L}_{A2H}}{\partial\mathbf{P}_{*i}} &= \frac{1}{2}\sum_{j=1}^{N}\left(\sigma(\mathbf{\Theta}_{ij}) - \mathbf{S}_{ij}^{(c)}\mathbf{Q}_{*j}\right)\\ &+ \frac{1}{2}\sum_{j=1}^{N}\lambda\mathbf{P}_{*j}\mathbf{S}_{i,j}^{(att)}\sigma(\phi_{i,j}\mathbf{S}_{i,j}^{(att)})(1-\sigma(\phi_{i,j}\mathbf{S}_{i,j}^{(att)}))\\ &+ 2\eta\left(\mathbf{P}_{*i}-\mathbf{B}_{*i}\right) + 2\eta\mathbf{P}_{*i}^{T}\mathbf{1}.\end{aligned} \tag{6}$$

Then, the gradient of weight in A2H Net can be calculated with $\frac{\mathcal{L}_{A2H}}{\partial P_{*i}}$ according to chain rule. The details of training A2H Net are shown in Algorithm 1. Using mini-batch Stochastic Gradient Descent algorithm, we fix the batch size to be 32. The initial learning rate is set as 10^{-3} and decreased by 0.01% for each iteration. We choose the hyperparameter λ and η in AgNet according to the results on validation set and find the best performances can be achieved with $\lambda = \eta = 1$. Therefore, we set $\lambda = \eta = 1$. The convergence curve of AgNet is shown in Fig.3, AgNet reaches a stable objective function value at the 30-th epoch, which indicates the efficiency of AgNet. Our neural network is implemented with TensorFlow library on an NVIDIA 1080ti GPU server.

Algorithm 1 Algorithm for training A2H Net.

Input:
The predicted visual attribute vectors $\hat{\mathbf{a}}^{(v)}$,
the predicted textual attribute vectors $\hat{\mathbf{a}}^{(t)}$,
label matrix $\mathbf{Y}$ and attribute matrix $\mathbf{A}$.

Output:
Parameters θ in the A2H Net and binary codes $\mathbf{B}$.

```
Initialization: Randomly initialize parameters $\theta_h$ of A2H Net, set mini-batch $M = 32$ and iteration number $l = \lfloor N/M \rfloor$.
Repeat
for $iter = 1, 2, ..., l$ do
    Randomly sample $M$ instances.
    Calculate category similarity $\mathbf{S}^{(c)}$.
    Calculate attribute similarity $\mathbf{S}^{(att)}$.
    Calculate $\mathbf{Q}$ and $\mathbf{P}$ by forward propagation, respectively.
    Get the corresponding binary code $\mathbf{B}$.
    Update the parameter $\theta$ by back propagation.
until a fixed number of iterations.
```

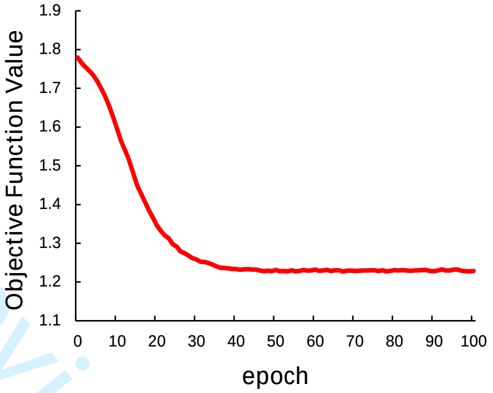


Fig. 3. Convergence curve of AgNet on AwA dataset.

IV. Experiment

In this section, we implement both the single-model and cross-modal zero-shot retrieval tasks, i.e., IBIR and TBIR, on three benchmark datasets. And we compare the proposed AgNet approach with several existing state-of-the-art methods to demonstrate its effectiveness.

A. Datasets

Animals with Attributes (AwA) [28]. AwA dataset consists of 30,475 images from 50 animal categories and 85 associated class-level attributes. It is a popular dataset for ZSL. We follow the standard seen/unseen split [28], where 40 categories with 24,295 images are taken as the seen domain and the remaining 10 categories with 6,180 images are adopted as the unseen domain.

TABLE III
Results on three benchmark datasets in Mean Average Precision (%) on CMZSH task. The best results are marked in bold.

Method	AwA					SUN					ImageNet				
	8bits	16bits	32bits	48bits	64bits	8bits	16bits	32bits	48bits	64bits	8bits	16bits	32bits	48bits	64bits
SCMH [18]	15.2	14.2	14.1	12.6	12.1	15.1	16.2	19.1	21.4	18.8	1.46	1.88	2.06	1.84	1.73
SMFH [20]	17.7	19.3	21.5	22.9	21.6	12.6	12.1	12.4	12.6	13.1	1.38	1.33	2.00	2.23	2.40
DCMH [21]	11.9	9.8	12.7	9.8	10.3	12.3	12.6	13.7	13.5	14.1	1.00	1.04	1.03	1.00	1.01
DCMH-G	12.4	10.1	11.8	12.3	13.5	13.4	12.8	12.7	13.2	14.0	1.19	1.27	1.38	1.62	1.57
FSH [17]	12.7	14.1	14.2	12.6	12.1	19.7	20.8	16.2	18.7	16.5	1.44	1.95	2.31	2.65	2.72
AgNet	**41.9**	**50.1**	**56.1**	**58.1**	**58.8**	**21.1**	**21.3**	**23.5**	**24.5**	**26.6**	**3.80**	**5.26**	**5.89**	**5.98**	**5.77**

SUN attribute [33]. It is another widely used dataset in ZSL, which consists of 717 scene categories annotated by 102 attributes. Each category has 20 images, and there are totally 14,340 images. Following [34], we utilize 707 categories as the seen domain and the other 10 categories as the unseen domain. It is worth noting that SUN dataset has image-level attribute annotations and its attribute is real-value. We utilize binary class-level attribute annotations in AgNet with the following consideration: 1) We view the attribute prediction as the classification problem, where labels used in deep neural network are usually in the form of one-hot code. 2) Binary label increases the sparsity of annotation. 3) The calculation of attribute similarity involves class-level category similarity, which makes it necessary to transfer the image-level annotation into the class-level. To get the class-level binary attribute annotations, we first calculate the average value for each category and then ceil the mean values to get the binary attributes.

ImageNet [35]. ImageNet is a large-scale image dataset organized according to the Word-Net [36] hierarchy. As no attribute is annotated for this dataset, in our experiment, we use AwA dataset as an auxiliary dataset to construct the training set. Specifically, after removing 10 similar categories shared by two datasets[1], we choose 40 categories with 21,832 images from AwA as seen domain and 100 animal categories with 129,622 images from ILSVRC2012 as the unseen domain.

B. Cross-Modal Zero-Shot Hashing

Under cross-modal zero-shot retrieval setting, *i.e.*, TBIR, the seen data are used for training the model. At the testing stage, the names of unseen categories are used as queries for retrieving images from the unseen domain.

Since the existing ZSH approaches cannot tackle the cross-modal retrieval task, we choose the CMH approaches for comparison. Four existing state-of-the-art CMH approaches are selected for comparison, where SCMH [18], SMFH [20], and DCMH [21] are three representative supervised CMH methods, while FSH [17] is an unsupervised CMH method. As the backbone net of visual modality in DCMH is a variant of AlexNet [39], we perform an additional experiment on DCMH by replacing it with GoogleNet for fair comparison, which is denoted as DCMH-G. For all comparative approaches, we use the codes provided by the authors. As DCMH is an end-to-end CMH method, we utilize raw images as input. The others adopt the same visual features as ours, that is, the GoogleNet features [32] fine-tuned in the training set. Besides, we use the word2vec features [23] as textual features for all methods.

We use the Mean Average Precision (mAP) to evaluate the performances of the proposed AgNet and the comparative approaches. To observe the performance under different code lengths, we set the code length with 8, 16, 32, 48 and 64 bits , respectively. From the results shown in Table III, we have the following observations: i) The proposed AgNet achieves the best performance on all three datasets consistently. All the comparative approaches have a relatively poor performance. This is mainly due to the reason that they are not designed for zero-shot settings, which leads to a worse generalization ability on the unseen domains. Specifically, it has a significant improvement on AwA dataset. For instance, the mAP performance of AgNet is 58.1% with 48 bits, which has a 35.2% absolute gain than that of the second best method SMFH. ii) The performances of AgNet on SUN dataset are inferior to those on AwA dataset. This is partly due to the fact that SUN is a fine-grained dataset in which there are few diversities in each category, making the learned hash codes be less discriminative. iii) AgNet also has a relatively small promotion on the large-scale ImageNet. Considering that the numbers of both the testing categories and instances in ImageNet are far more than those in AwA and SUN datasets, the improvements are still impressive. iv) The mAP performances of AgNet are positively related to code length in most situations, which indicates that the discriminative ability of hash codes increases with the growth of code length. By contrast, the mAP performances of comparative methods are unstable and without such a property. In a word, the experimental results clearly demonstrate the superiority of the proposed AgNet approach in CMZSH task.

C. Single-Modal Zero-Shot Hashing

The existing ZSH methods mainly focus on single-modal retrieval, *i.e.*, the query set and retrieval set are both constructed with the images. To evaluate the generalization of AgNet, we also implement AgNet in the single-modal ZSH task. As the scale of the unseen domain in SUN is insufficient to evaluate the performance of image retrieval task, we just implement single-model retrieval on AwA and ImageNet datasets.

Following [12] and [13], we randomly choose 10,000 instances from seen domain to construct the training set. As for testing, we randomly select 1,000 images from the unseen domain as the query set. The remaining unseen images and all seen domain images form the retrieval set.

[1] We eliminate 10 categories (*i.e.*, dalmatian, collie, german shepherd, chihuahua, persian cat, siamese cat, bobcat, horse, deer, sheep) from AwA to construct the seen domain.

We select the following state-of-the-art hashing methods as the baselines. IMH [37] and ITQ [5] are two representative unsupervised hashing methods, SDH [6], TSK [12] and SitNet [13] are three supervised hashing methods. In addition, TSK and SitNet are specially designed for zero-shot retrieval. The mAP and Precision within Hamming radius 2 are adopted as the evaluation metrics in this task. For all comparative approaches, we utilize GoogleNet features fine-tuned in the training set as the visual features. Following [13], we set the code length to be 8, 16, 32, and 48 bits, respectively.

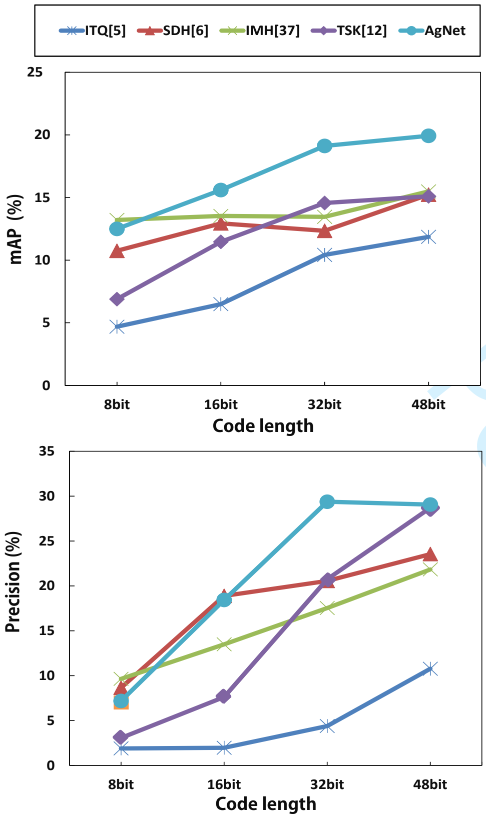


Fig. 4. Performances (mAP and Precision) of different methods for single-modal zero-shot hashing task on AwA dataset.

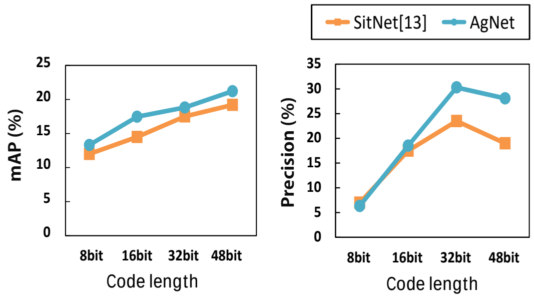


Fig. 5. Performances (mAP and Precision) of SitNet and AgNet for single-modal zero-shot hashing task on AwA dataset.

1) Experimental results on AwA: All the comparative approaches are implemented by ourselves with the code provided by the authors, except for SitNet. The results of SitNet are directly cited from the original paper [13]. It should be noted that the split of seen and unseen domain in SitNet has a slight difference with ours. SitNet randomly chooses 10 categories as unseen domain, while we follow the standard split [28] in this work. We follow this setting to make our work repeatable. For fair comparison, we randomly split the seen and unseen domain following the SitNet and individually compare AgNet and SitNet.

The performances of AgNet and the comparative methods on AwA dataset are reported in Fig. 4 and Fig. 5, respectively. As we can see, the proposed AgNet achieves the best mAP performance in most cases. For example, AgNet gains 18.8% on 32 bits, which has an improvement against SitNet by 7.4% in the same code length. Besides, the mAP performances of AgNet keep improving with the increase of code length, which is similar to the phenomenon in the cross-modal retrieval task. As for Precision, AgNet exceeds all comparative methods in the code length of 32 and 48 bits, and only achieves a slightly inferior performance on 8 and 16 bits. Moreover, there is a slight drop from 32 bits to 48 bits in the precision performance of AgNet, indicating that we need to choose a suitable code length to guarantee the retrieval performance.

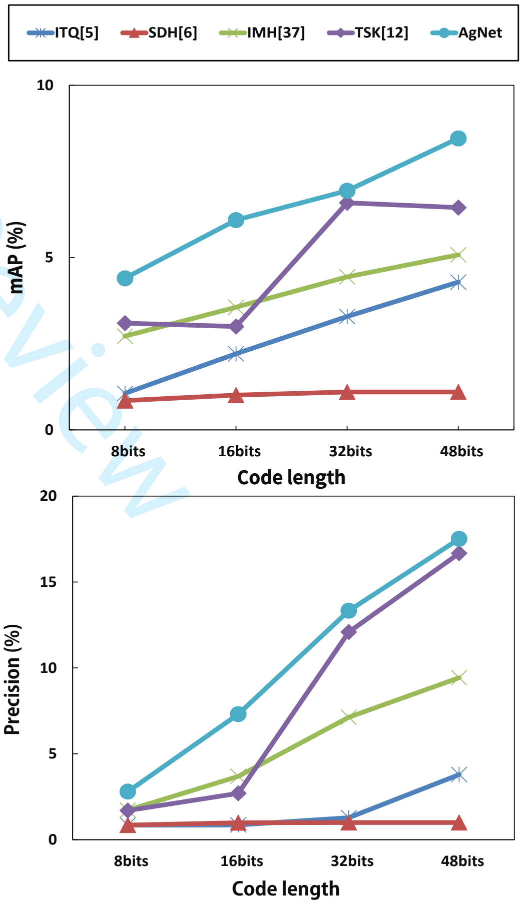


Fig. 6. Performances (mAP and Precision) of different methods for single-modal zero-shot hashing task on ImageNet dataset.

2) Experimental results on ImageNet: The comparative experiments are reported in Fig. 6. Note that SitNet [13] is not selected for comparison in this dataset since its experimental setting is different from ours. It can be observed that AgNet

TABLE IV
Performances (mAP / Precision) of AgNet and AgNet-vis on ImageNet dataset on SMZSH task. The best results are marked in bold.

Method	8bits	16bits	32bits	48bits
AgNet-vis	4.33 / 2.41	5.66 / 5.11	6.92 / 10.7	8.22 / 15.71
AgNet	**4.40 / 2.80**	**6.09 / 7.31**	**6.94 / 13.33**	**8.46 / 17.51**

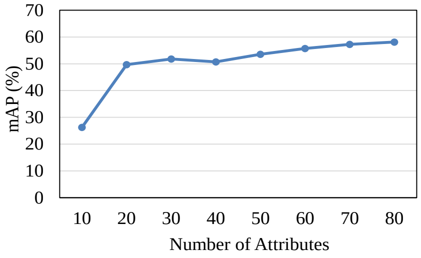


Fig. 7. Performances of AgNet with different number of attributes on AwA dataset.

outperforms all comparative methods with significant margins in all code lengths. Besides, the performances of unsupervised methods surpass those of the conventional supervised method, *i.e.*, SDH. Without using the supervision information, unsupervised methods exploit the inherent property of visual representations to generate hash codes and avoid suffering from the misleading of supervision information of seen categories in zero-shot setting. However, by utilizing semantic information as the transferable supervision information, AgNet and TSK mitigate the influence of zero-shot problem and outperform both unsupervised and conventional supervised hashing methods on ImageNet dataset.

3) The contribution of T2A Net for SMZSH task: It can be noticed that T2A net of AgNet is adopted on SMZSH task.To evaluate the effect of T2A Net, we implement experiment with a visual version of AgNet, which is denoted as AgNet-vis, on ImageNet dataset. The main different between AgNet and AgNet-vis is that Agnet-vis removes the T2A Net. In addition, to maintain the category similarity, AgNet-vis replaces $\mathbf{\Theta}_{ij} = \frac{1}{2}\mathbf{P}_{*i}{}^{T}\mathbf{Q}_{*j}$ with $\phi_{ij} = \frac{1}{2}\mathbf{P}_{*i}^{T}\mathbf{P}_{*j}$ in category similarity loss function. The experimental results on ImageNet is shown in Table IV. It can be observed that AgNet outperforms AgNet-vis in all code lengths consistently, which demonstrates that the additional textual information is beneficial to generation of effective visual hash codes.

D. *Effects of Attributes*

As our algorithm is an attribute-based method, the performance of the learned attribute space will affect the discriminative ability of hash codes. In this part, we implement some experiments to analyze the impact of the attribute space, including the scale of attribute space and the attribute prediction accuracy to the final performances.

To evaluate the influence of attribute space scale, we vary the number of attributes from 10 to 80 with the interval of 10. In consideration of the difference on attributes, we report the average performance of 5 trials for each number by fixing the code length as 64 bits. The curve of CMZSH in terms of mAP is shown in Fig. 7.

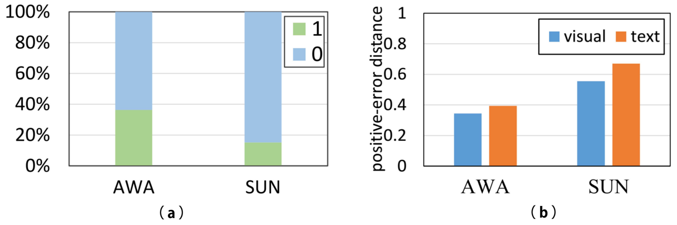


Fig. 8. (a)The distribution of binary attribute tags on AwA and SUN. (b)The results of positive-error distance on AwA and SUN.

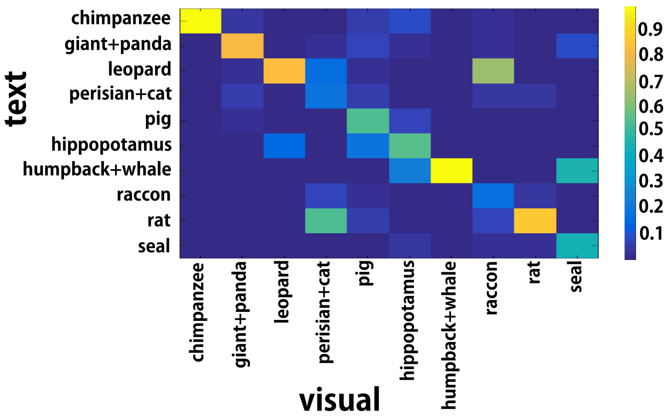


Fig. 9. Confusion matrix of AgNet on AwA, where the columns are the categories that visual hash codes belong to and the rows are the categories of textual hash codes that visual hash codes are close to.

It can be observed that the mAP performance increases with the growth of attribute number. Specifically, there is a giant leap when attribute number changes from 10 to 20. It indicates that more attributes are required to guarantee the discriminative ability. Besides, when the amount of attribute is large enough, the increasing scope turns to saturation.

In addition, the attribute prediction accuracy also plays a significant role in the performance of AgNet. The previous experiment in cross-modal task has demonstrated that the performances of AgNet on SUN are inferior to those on AwA. The underlying reason may be that the attribute prediction accuracy on SUN is inferior to those on AwA. Therefore, we analyze the attribute prediction accuracy on both datasets.

According to the distribution of binary attribute tags on AwA and SUN, as is shown in Fig. 8(a), it can be easily noticed that the tags of AwA and SUN are biased to 0. Therefore, we propose to utilize the positive-error distance(PED) to evaluate the prediction accuracy, which is defined as:

$$\mathcal{D} = \frac{\sum_{i}^{N}\sum_{j}^{d}\mathbf{A}_{ji}\left|\mathbf{A}_{ji} - \hat{\mathbf{A}}_{ji}\right|}{\sum_{i}^{N}\sum_{j}^{d}\mathbf{A}_{ji}}, \tag{7}$$

where $\hat{\mathbf{A}}_{*i}$ denotes the predicted attribute vector and $\mathbf{A}_{*i}$ denotes the ground-truth attribute vector, d is the dimensionality of $\mathbf{A}_{*i}$ and N is the number of instances. Using Eq. (7), the distances between $\mathbf{A}_{*i}$ and $\hat{\mathbf{A}}_{*i}$ are calculated when $\mathbf{A}_{ji} = 1$.

The results are reported in Fig. 8(b), which demonstrate that the attribute prediction on AWA is closer to the ground

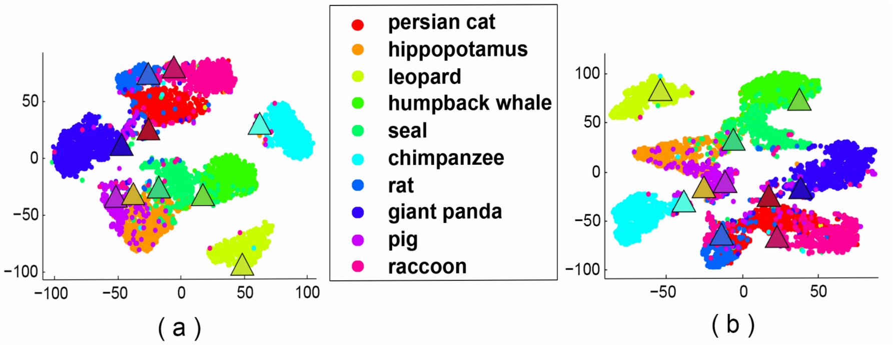


Fig. 10. t-SNE visualization of unseen instances on AWA dataset. Points denote visual representations and triangles denote text representations. (a) The visualization of attribute predictions. (b) The visualization of outputs from the last layer in A2H Net.

truth than that on SUN. The PED of visual modality and textual modality on SUN are larger than that on AwA in 70.1% and 61.4%, respectively. Thereby, the attribute prediction on AwA is more discriminable than that on SUN, which further interprets the better performance of AgNet on AwA than those on SUN.

E. Visualization

To further evaluate the performance of AgNet in each category, taking AwA dataset for example, we utilize confusion matrix to visualize the neighbor relationship between textual hash codes and visual hash codes of AgNet. We fix the code length to 64 bits. The result is shown in Fig. 9, where each column denotes the categories that visual instances belong to, and each row is the categories of textual instances that visual instances are close to. It can be observed that most instances are concentrated in the diagonal line, which indicates that visual instances are close to the text instance with the same category in most situations. However, there still exists some confusions in some categories. Take “seal” as example, about 40% of visual instances are close to “humpback whale”. The main underlying reason is that both categories are marine animal with a lot of similar attributes, which misguides the model to generate the similar hash codes for both categories. This means that the performance of AgNet in similar categories should be further improved in future.

In addition, the hash codes need to preserve the neighbor relationship of the original features. As for AgNet, we use A2H Net to generate hash codes from both the textual and visual modalities. In this part, we use t-SNE [38] to visualize the performance of A2H Net on the unseen domain. Instead of adopting the binary codes that are difficult to generate effective cluster with t-SNE, we utilize attribute predictions and outputs from the last layer in A2H Net as the inputs for t-SNE. As is illustrated in Fig. 10, we can observe that the similarity relationship in attribute space has been well preserved in the hash space. For instance, the visual and textual instances from “leopard” matain the same relationships with other classificatory instances in attribute and hash space.

V. Conclusion

In this paper, we have proposed a deep hashing neural network to address the cross-modal zero-shot retrieval problem. It aligns different modal data into a more high-level semantic space, *i.e.*, attribute space. Besides, category similarity is utilized to construct the relationships between different modalities while attribute similarity is introduced to regularize the distance of similar categories in single modality. Experimental results on both cross-modal and single-modal retrieval tasks have demonstrated the superiority of the proposed approach.

In the future, as the acquisition of attribute annotation requires prior knowledge, we plan to exploit some other semantic information to formulate the common space, *e.g.*, click-through data. We will also apply Hashlayer described in [40] to control the quantization error. In addition, we will exploit generative methods, *e.g.*, GAN and VAE, to establish more robust embedding in zero-shot hashing.

References

[1] L. Liu, and L. Shao, “Sequential compact code learning for unsupervised image hashing,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 12, pp. 2526-2536, 2016.

[2] L. Liu, M. Yu, and L. Shao, “Latent structure preserving Hashing,” *Int. Jou. of Comput. Vis.*, vol. 122, no. 3, pp. 439-457, 2017.

[3] L. Lin, G. Ding, J. Han, and J. Wang, “Cross-view retrieval via probability-based semantics-preserving hashing,” *IEEE Trans. on Cybern.*, vol. 47, no. 12, pp. 4342-4355, 2017.

[4] F. Shen, C. Shen, Q. Shi, A. Hengel, Z. Tang, and H. Shen, “Hashing on nonlinear manifolds,” *IEEE Trans. on Image Process.*, vol. 24, no. 6, pp. 1839-1851, 2015.

[5] Y. Gong, S. Lazebnik, A. Gordo, and F. Perronnin, “Iterative Quantization: a procrustean approach to learning binary codes for large-scale image retrieval,” *IEEE Trans. on Pattern Anal. Mach. Intell.*, vol. 35, no. 12, pp. 2916-2929, 2013.

[6] F. Shen, C. Shen, W. Liu, and H. Shen, “Supervised discrete hashing,” in *Proc. IEEE Conf. on Comput. Vis. Pattern Recognit.*, Boston, USA, June 2015, pp. 37-45.

[7] W. Liu, J. Wang, R. Ji, Y. Jiang, and S. Chang, “Supervised hashing with kernels,” in *Proc. IEEE Conf. on Comput. Vis. Pattern Recognit.*, Providence, USA, June 2012, pp. 2074-2081.

[8] C. Lampert, H. Nickisch, and S. Harmeling, “Attribute-based classification for zero-shot visual object categorization,” *IEEE Trans. on Pattern Anal. Mach. Intell.*, vol. 36, no. 3, pp. 453-465, 2014.

[9] Y. Fu, T. Hospedales, T. Xiang, and S. Gong, “Learning multimodal latent attributes,” *IEEE Trans. on Pattern Anal. Mach. Intell.*, vol. 36, no. 2, pp. 303-316, 2014.

[10] L. Yang, L. Liu, L. Shao, F. Shen, G. Ding, and J. Han, “From zero-shot learning to conventional supervised classification: unseen visual data synthesis,” in *Proc. IEEE Conf. on Comput. Vis. Pattern Recognit.*, Honolulu, USA, July 2017, pp. 6165-6174.

[11] Z. Ji, Y. Yu, Y. Pang, J. Guo, and Z. Zhang, “Manifold regularized cross-modal embedding for zero-shot learning,” *Inf. Sci.*, pp. 48-58, 2017.

[12] Y. Yang, Y. Luo, L. Chen, F. Shen, J. Shao, and H. Shen, “Zero-shot hashing via transferring supervised knowledge,” in *Proc. ACM Conf. on Multimedia*, Amsterdam, Netherlands, Oct. 2016, pp. 1286-1295, 2016.

[13] Y. Guo, G. Ding, J. Han, and Y. Guo, "SitNet: discrete similarity transfer network for zero-shot hashing," in *Int. Joint Conf. on Art. Intell.*, Melbourne, Australia, Aug. 2017, pp. 1767-1773.

[14] J. Wang, W. Liu, S. Kumar, and S. Chang, "Learning to hash for indexing big data - a survey," *Proc. of the IEEE*, vol. 104, no. 1, pp. 34-57, 2016.

[15] Y. Fu, T. Xiang, Y. Jiang, X. Xue, L. Sigal, and S. Gong, "Recent advances in zero-shot recognition: toward data-efficient understanding of visual content," *IEEE Signal Process. Maga.*, vol. 35, no. 1, pp. 112-125, 2018.

[16] G. Ding, Y. Guo, J. Zhou, and Y. Gao, "Large-scale cross-modality search via collective matrix factorization hashing," *IEEE Trans. on Image Process.*, vol. 25, no. 11, pp. 5427-5440, 2016.

[17] H. Liu, R. Ji, Y. Wu, F. Huang, and B. Zhang, " Cross-modality binary code learning via fusion similarity hashing," in *Proc. IEEE Conf. on Comput. Vis. Pattern Recognit.*, Honolulu, USA, July 2017, pp. 6345-6353.

[18] D. Zhang, and W. Li, "Large-scale supervised multimodal hashing with semantic correlation maximization," in *AAAI Conf. Art. Intell.*, Qubec, Canada, July 2014, pp. 2177-2183.

[19] Z. Lin, G. Ding, M. Hu, and J. Wang, "Semantics-preserving hashing for cross-view retrieval," in *Proc. IEEE Conf. on Comput. Vis. Pattern Recognit.*, Boston, USA, June 2015, pp. 3864-3872.

[20] H. Liu, R. Ji, Y. Wu, and G. Hua, "Supervised matrix factorization for cross-modality hashing," in *Int. Joint Conf. on Art. Intell.*, New York, USA, Jan. 2016, pp. 1767-1773

[21] Q. Jiang, and W. Li, "Deep cross-modal hashing," in *Proc. IEEE Conf. on Comput. Vis. Pattern Recognit.*, Honolulu, USA, July 2017, pp. 3270-3278.

[22] E. Yang, C. Deng, W. Liu, X. Liu, D. Tao, and X. Gao, "Pairwise relationship guided deep hashing for cross-modal retrieval," in *AAAI Conf. Art. Intell.*, San Francisco, USA, Feb. 2017, pp. 1618-1625.

[23] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," Advances in *Neural Inf. Process. Syst.*, Nevada, USA, Dec. 2013, pp. 3111-3119.

[24] Y. Xu, Y. Yang, F. Shen, X. Xu, Y. Zhou, and H. Shen, "Attribute hashing for zero-shot image retrieval," in *IEEE Int. Conf. on Multimedia and Expo*, Hong Kong, China, July 2017, pp. 133-138.

[25] Y. Fu, M. Hospedales, Timothy, T. Xiang, and S. Gong, "Transductive multi-view zero-shot learning," *IEEE Trans. on Pattern Anal. Mach. Intell.*, vol. 37, no. 11, pp. 2332-2345, 2015.

[26] Y. Yu, Z. Ji, J. Guo, and Y. Pang, "Transductive zero-shot learning with adaptive structural embedding," *IEEE Trans. Neural Netw. Learn. Syst.*, pp. 1-12, 2017.

[27] H. Lai, and Y. Pan, "Transductive zero-shot hashing via coarse-tofine similarity mining," arXiv:1711.02856, 2017.

[28] C. Lampert, H. Nickisch, and S. Harmeling, "Learning to detect unseen object classes by between-class attribute transfer," in *Proc. IEEE Conf. on Comput. Vis. Pattern Recognit.*, Florida, USA, June 2009, pp. 951-958.

[29] Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid, " Label-embedding for attribute-based classification," in *Proc. IEEE Conf. on Comput. Vis. Pattern Recognit.*, Portland, USA, June 2013, pp. 819-826.

[30] Z. Zhang, and Saligrama, "Zero-shot recognition via structured prediction," in *Eur. Conf. on Comput. Vis.*, Amsterdam, Netherlands, Oct. 2016, pp. 533-48.

[31] M. Rohrbach, M. Stark, G. Szarvas, I. Gurevych, and B. Schiele, " What helps where and why? Semantic relatedness for knowledge transfer," in *Proc. IEEE Conf. on Comput. Vis. Pattern Recognit.*, San Francisco, USA, June 2010, pp. 910-917.

[32] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, " Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. on Comput. Vis. Pattern Recognit.*, Las Vegas, USA, June 2016, pp. 2818-2826.

[33] G. Patterson, and J. Hays, "SUN attribute database: Discovering, annotating, and recognizing scene attributes," in *Proc. IEEE Conf. on Comput. Vis. Pattern Recognit.*, Providence, USA, June 2012, pp. 2751-2758.

[34] D. Jayaraman, and K. Grauman, "Zero-shot recognition with unreliable attributes," Advances in *Neural Inf. Process. Syst.*, Montral Canada, Dec. 2014, pp. 3464-3472.

[35] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and F.F. Li, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. on Comput. Vis. Pattern Recognit.*, Florida, USA, June 2009, pp. 248-255.

[36] G. Miller, "Wordnet: a lexical database for english," in *Comm. of the ACM*, Nov. 1995, vol. 38, no. 11, pp. 39-41.

[37] F. Shen, C. Shen, Q. Shi, A. Hengel, and Z. Tang, " Inductive hashing on manifolds," in *Proc. IEEE Conf. on Comput. Vis. Pattern Recognit.*, Portland, USA, June 2013, pp. 1562-1569.

[38] L. Maaten, and G. Hinton, "Visualizing data using t-sne," *J. Mach. Learn. Res.*, vol. 9, pp.2579-2605, 2008.

[39] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the Devil in the Details: Delving Deep into Convolutional Nets," in *Proc. of British Mach. Vis. Conf.*, Nottingham, UK, Sep. 2014.

[40] Z. Cao, M. Long, J. Wang, and P. Yu, "HashNet: Deep Learning to Hash by Continuation," in *Int. Conf. Comput. Visi.*, Venice, Italy, Oct. 2017, pp. 5609-5618.