Retrieval Across Optical and SAR Images with Deep Neural Network

Yifan Zhang, Wengang Zhou(✉), and Houqiang Li

CAS Key Laboratory of Technology in Geo-spatial, Information Processing and Application System, Department of Electronic Engineering and Information Science, University of Science and Technology of China, Hefei 230027, China
zyf1@mail.ustc.edu.cn, {zhwg,lihq}@ustc.edu.cn

Abstract. In this paper, we are dedicated to the cross-modal image retrieval between optical images and synthetic aperture radar (SAR) images. This cross-modal retrieval is a challenging task due to the different imaging mechanisms and huge heterogeneity gap. Here, we design a two-stream fully convolutional network to tackle this issue. The network maps the optical and SAR images to a common feature space for comparison. For different modal images, the comparable features are obtained by feeding them into the corresponding branch. Each branch fuses two types of features in a weighted manner. These two kinds of features root in the pooling features of VGG16 at different depths, but are refined by the well-designed channels-aggregated convolution (CAC) operation as well as semi-average pooling (SAP) operation. In order to get a better model, an extensible training approach is proposed. The training of the model is from the local to the whole. Besides, we collect an optical/SAR image retrieval (OSR) dataset. Comprehensive experiments on this dataset demonstrate the effectiveness of our proposed method.

Keywords: Retrieval · Cross-modal
Convolutional Neural Network (CNN) · Feature fusion

1 Introduction

With the development of sensor technology, there is an increasing number of remote sensing data with diversified modalities. Through integrating multimodal data which represent the same content, many tasks [12,23,26] have been reinvestigated. Therefore, how to find relevant data between different modalities is an active topic. Here, we focus on the retrieval across optical and SAR images. These two modal images cover the same ground scene but are vastly different. Hence, this is also a cross-modal problem. If we successfully search the matching images, it will be useful for subsequent tasks, such as height estimation [27].

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-030-00776-8_36) contains supplementary material, which is available to authorized users.

R. Hong et al. (Eds.): PCM 2018, LNCS 11164, pp. 392–402, 2018.
https://doi.org/10.1007/978-3-030-00776-8_36

In recent years, many cross-modal retrieval methods have been developed. Fukui *et al.* [9] employ matching correlation analysis for retrieval between image and tag. In [21], Sharma *et al.* use partial least squares to handle the cross-modal face recognition with huge variation in resolution. However, most of these methods not only separate feature extraction and subspace generation [24], but also need external data preprocessing. Enlightened by the deep learning techniques, Li *et al.* [15] employ the deep features to improve the retrieval performance across the street and shop domains. Qi *et al.* [19] use siamese network to better complete the sketch-based cross-modal image retrieval. Moreover, the deep models are also effective in remote sensing field. Zhong *et al.* [29] use deep belief networks to classify hyperspectral images. In [6], CNN is used for object detection in optical images. Nevertheless, in remote sensing community, the cross-modal retrieval using deep networks has been rarely explored.

Motivated by the above observations and the fact that the deep features are usually more discriminative [25], a cross-modal remote sensing image retrieval method is proposed. The retrieval method is based on a deep two-stream network, where each branch is the same fully convolutional network to process a specific domain, yet their weights are different. For each branch, we aim to learn proper weights to map the different modal images to a common feature space [11], in which relevant data are closer to each other. In our task, we want to learn a correct deep model so that using an optical image can find the matching SAR images, which cover the same ground scene with optical image.

In this paper, the proposed deep model simultaneously utilizes the visual appearance and powerful semantic representation by fusing features from different layers, so we name it as Double-Feature Convolutional Neural Network (DFCNN). The framework is shown in Fig. 1. Especially, the channels-aggregated convolution and semi-average pooling operations are applied in DFCNN to make full use of the semantic information and reduce the feature dimension. Inspired by the coarse-to-fine training strategy, an extensible training method is proposed. We first train each branch of the model by other tasks (*e.g.,* semantic segmentation task and autoencoder model), and then fine-tune on the pre-trained model. Finally, we collect an optical/SAR image retrieval (OSR) dataset to study this task, and the dataset will be released to the public. To the best of our knowledge, the dataset is proposed for the first time. The experiments on this dataset demonstrate the effectiveness of our retrieval method.

2 Our Approach

In this paper, we focus on the cross-modal retrieval across optical and SAR images. Inspired by the deep learning methods, a two-stream deep network is designed. The objective is to use this network to learn effective feature representations, which make the distances between matched pairs smaller than the mismatched pairs. By training with contrastive loss [7], the network can learn proper weights to map optical and SAR images into a common feature space. Each branch handles a modal image here. For the two branches, the learned features which are in the common space, can be used for our cross-modal retrieval.

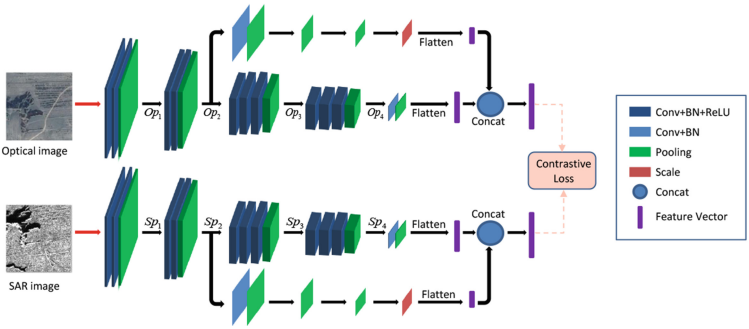

Fig. 1. The proposed two-steam fully convolutional network. Each branch contains ten convolutional layers from VGG16, and the feature fusion part. The fused two types of features are refined by the channels-aggregated convolution as well as semi-average pooling operation, which is explained well in Sect. 2.1. We train the network with contrastive loss which can force the features from two branches to a common space.

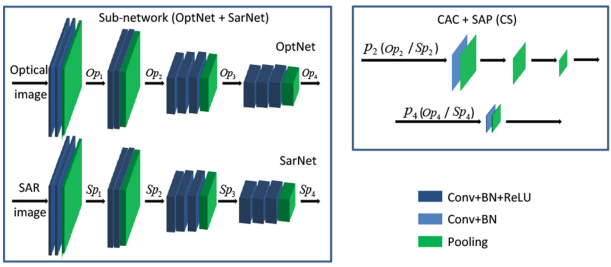

Fig. 2. Some important modules in our proposed network, *i.e.,* DFCNN. In the Sub-network, there are two branches, named as OptNet in optical branch and SarNet in SAR branch, respectively (left). The well-designed channels-aggregated convolution (CAC) and semi-average pooling (SAP) operations on p_2 and p_4 (right).

In this section, we first introduce the network architecture of DFCNN and the effective training method in detail. Then, the retrieval method based on the trained deep model is proposed.

2.1 Network Structure

The proposed network architecture, *i.e.,* DFCNN is depicted in Fig. 1, and some important modules are also illustrated in Fig. 2 for clarity. The designed deep model mainly contains two parts. One is the Sub-network shown in Fig. 2, and the other fuses the features on two levels. In the Sub-network which is the backbone of DFCNN, there are two branches to handle optical and SAR images separately, named as OptNet and SarNet, respectively. Each branch only uses the first four blocks of convolutions of VGG16 [22], *i.e.,* ten convolutional layers. The reasons are as follows. (1) The fully convolutional network can better learn contextual

information. (2) The lighter network already has the adequate ability to solve our problem and reduces the difficulty in training.

Inspired by the skip connections [4], it is feasible to improve the search performance by fusing the low-level and high-level features [13]. In general, the deeper the network is, the higher level of abstraction the features provide. In our work, we choose to fuse the features from the second and fourth pooling layers in Subnetwork, *i.e.,* p_2 and p_4 in Fig. 2. Specifically, the Op_2 and Op_4 (Sp_2 and Sp_4) are fused in optical (SAR) branch. The p_4 is a necessary choice for bridging the semantic gap because it is more discriminative. Besides, p_2 is more helpful than p_1 because it carries more abstract knowledge. Moreover, compared to the p_4 and p_3, p_4 and p_2 has fewer semantic conflicts because of the farther distance. However, the original p_4 and p_2 are rough with very high dimensions. Enlightened by the fact that human understands the world by holistic perceptions from various views and a reasonable spatio-temporal range can bring right judgment [8], we design the corresponding channels-aggregated convolution (CAC) and semi-average pooling (SAP) schemes to refine features.

As illustrated in Fig. 2, the CAC and SAP operations are performed on p_2 and p_4. The CAC operation aggregates the scattered features on different feature maps into global features. Assume that the CAC operation is performed on the feature maps with C channels, one CAC feature f (*i.e.,* one global feature) is denoted as follows:

$$f = \sum_{i=1}^{C} W_i \bullet Conv_i(x_i), \tag{1}$$

where $Conv_i(x_i)$ indicates using the i-th kernel convolves the i-th feature map. W_i is the corresponding weight to balance the contribution of each feature map. In this paper, the CAC operation is implemented by applying a convolutional layer, whose number of output channels corresponds to the number of generated CAC features.

Since the downsampling can eliminate redundant features and make better use of the contextual information, we choose pooling operation to enhance semantic representation. Babenko *et al.* [2] point out a simple global descriptor based on sum pooling aggregation performs well on standard retrieval datasets. However, such descriptor to aggregate the raw features ignores the influence of spatial resolution. Based on the sensing mechanism of human, we advocate the semi-average pooling (SAP) operation which can retain more beneficial information. The SAP operation downsamples the feature maps to a suitable resolution rather than one value. Empirically, the appropriate resolution is considered as half of the size of p_4. In our experiments, the CAC and SAP operations have been demonstrated based on p_4, respectively.

The detailed configuration of DFCNN can be found in the supplementary material. For the semi-average pooling operation, the avg-pooling is selected instead of the max-pooling to reduce information loss [2]. Meanwhile, the ReLU non-linearity is not used after Batch Normalization (BN) in the channels-aggregated convolution (CAC), because the negative values also contain sig-

nificant contents. Besides, two types of features are fused in a weighted manner to form the last feature. The weighting factor is learned in the network by adding a scale layer.

2.2 Training Approach

Since a proper initialization allows the network to converge quickly and correctly, a bottom-up training approach is devised. First, we adopt the encoder-decoder architecture about optical image (OED) to get the pre-trained OptNet in Fig. 2. The OED comes from the SegNet [3], but there are some differences. Only the first ten convolutional layers of SegNet are used in encoder part, and only four upsampling layers are used in decoder part. Besides, the softmax function is replaced by the sigmoid function. Because extracting buildings is a simplified semantic segmentation task, which just has two categories. The weights of OED model are initialized with the weights of the model[1]. Then, the OED model is trained on the Massachusetts Buildings Dataset [17] because of the similarity of the optical data. We optimize the OED model by minimizing the cross-entropy loss.

The encoder-decoder architecture about SAR image (SED) is similar to OED but discards the sigmoid layer. The input of SED is the SAR images in the training set, and the training target is to reconstruct the SAR images, like the autoencoder (AE) model [20]. Hence, the Euclidean distance loss is used to optimize the SED model. The weights of SED model are initialized by MSRAFiller [10]. When the training of OED and SED model is completed, their encoder parts can be used as the pre-trained OptNet and SarNet, respectively.

In the next training steps, all the networks are optimized with contrastive loss. And the input of networks is the image pairs on the OSR training set. First, we refine the Sub-network based on the pre-trained OptNet and SarNet. Next, for each branch, the channels-aggregated convolution and semi-average pooling operations are executed on features of two levels, and then we fuse the two types of features. Lastly, the complete DFCNN is fine-tuned with the fixed weights of pre-trained Sub-network.

2.3 End-to-end Retrieval

When the DFCNN is trained to convergence, we can use this DFCNN model to map optical and SAR images into the common feature space, in which relevant images are closer to each other. Here, each modal images are analyzed on the corresponding branch to get the compared features. To complete the cross-modal search, we only need to compare the similarity of output features between the query and database. The similarity is measured with Euclidean distance here. Hence, the retrieval across optical and SAR images is now performed in an end-to-end manner. In our experiments, both the optical image query and SAR image query are studied. The results are quantitatively evaluated with mAP and *Recall-K*. *Recall-K* denotes the *Recall* when the sorted index ranks top-K.

[1] http://mi.eng.cam.ac.uk/~agk34/resources/SegNet/segnet_pascal.caffemodel.

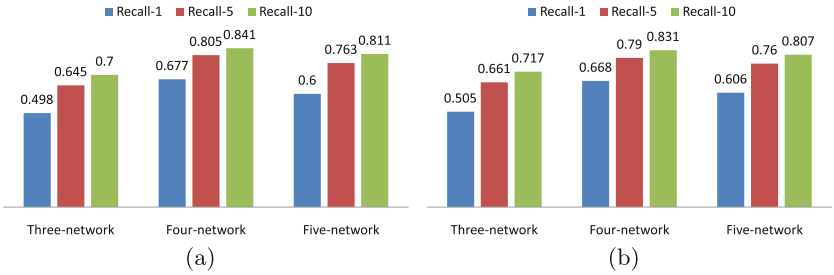

Fig. 3. The results of three different Sub-network. (a) Optical image as the query. (b) SAR image as the query.

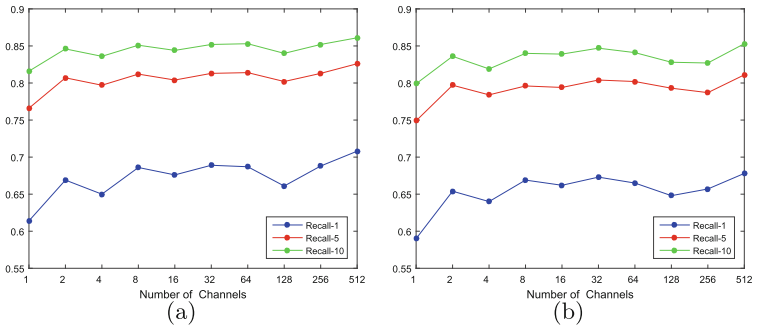

Fig. 4. The results when we produce different numbers of CAC features. Here are 10 different cases in all. (a) Optical image as the query. (b) SAR image as the query.

3 Experiments

3.1 Dataset Description

According to the Homography matrix, we first use the given 29 X-SAR grayscale images and the rough matched optical images from BIGEMAP[2] to get the ripe matched image pairs with the size of 8192×10240. Then, the slider operation is performed on each image pair with the stride of 90 pixels to generate our OSR dataset. Some screening conditions (*e.g.,* the number of black dots) are used to remove the meaningless data here. Finally, the dataset includes 46474 optical image patches with the size of 227×227. Besides, there are 5 positive pairs and 10 negative pairs for each optical image. In our setup, the dataset is divided into training set, validation set and test set in a proportion of $7 : 2 : 1$. The detailed construction of the dataset can be found in the supplementary material.

[2] http://www.bigemap.com/.

Table 1. Different groups when the map resolution is 8×8.

Feature Fusion	optical query	SAR query
	Recall-1	*Recall-1*
(p_4+2conv+1pool)ⓒ(p_1+1conv+4pool)	0.672	0.650
(p_4+2conv+1pool)ⓒ(p_2+1conv+3pool)	**0.675**	**0.671**
(p_4+2conv+1pool)ⓒ(p_3+1conv+2pool)	0.654	0.651

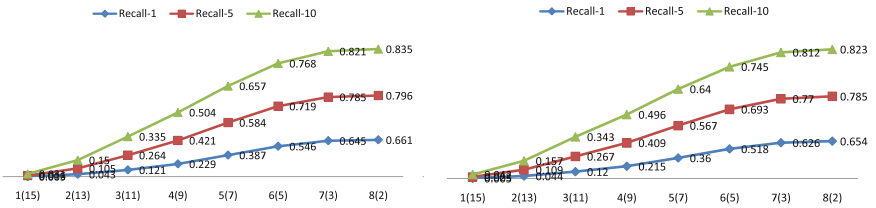

Fig. 5. The results of different resolutions from 1×1 to 8×8. The 8(2) in axis denotes that using 2×2 pixel window downsamples the feature maps to 8×8 resolution and so forth. (a) Optical image as the query. (b) SAR image as the query.

3.2 Experiments and Discussion

The networks are implemented in Caffe [14] and optimized with SGD algorithm [5] with momentum. We set the initial learning rate to 10^{-3}, momentum to 0.9 and batch size to 5. Following the proposed training method, the OED and SED model are first trained about 180k and 150k iterations to get the pre-trained OptNet and SarNet, respectively. And then, the Sub-network and DFCNN are trained about 800k and 160k iterations, respectively. Besides, we conduct the ablation studies and the comparative experiments with baseline methods.

Exploration of Sub-network, CAC and SAP. Here, we compare the Sub-network with different numbers of convolution blocks (from 3 to 5) of VGG16. Each branch of Four-network uses ten convolutional layers, *i.e.,* the first four convolution blocks of VGG16. As shown in Fig. 3, the Four-network which is used by our DFCNN, works best with different metrics. The reasons are as follows. (1) Each branch of Three-network only has seven convolutional layers and can not learn more semantic representation. (2) Each branch of Five-network has 13 convolutional layers, which makes the training more difficult.

From the Four-network, we can obtain the output feature p_4 (including Op_4 and Sp_4), each of which has 512 feature maps with 15×15 resolution. Here, the same channels-aggregated convolution (CAC) operation is performed on Op_4 and Sp_4, and then we compare the output CAC features. The number of channels of generated CAC features is determined by one convolutional layer, *i.e.,* how many filters are used. The results are shown in Fig. 4. In most cases, there is no

significant performance degradation. However, the result becomes worse when only one CAC feature is used. It is mainly because the sole feature contains incomplete information. By adding an avg-pooling layer with different kernel size on p_4 (Op_4 and Sp_4), we can get the features of different resolutions. For the avg-pooling layers, the stride is always 2 with padding 0. As shown in Fig. 5, the resolution is crucial. The result only drops by about 1 percent when the resolution changes from original 15×15 on p_4 to 8×8. However, the performance degrades a lot with other sizes. The motivation behind the semi-average pooling (SAP) is that the suitable spatial information is important. Hence, this result provides the basis for our choice of SAP rather than global pooling.

Evaluation on Feature Fusion, Weighting and Training Method. In this paper, if two channels-aggregated convolution (CAC) features are generated, *i.e.,* two filters are used to convolve input features, it is denoted as 2conv and so forth. Therefore, p_4+2conv+2pool denotes that we perform 2conv, and followed by twice avg-pooling on p_4 (Op_4 and Sp_4). The parameters of convolutional layer in CAC operation and pooling layer refer to the configuration of DFCNN. Besides, $A©B$ denotes the concatenation between feature A and B. For each branch, we carry out the corresponding feature fusion and then compare the output features of two branches. As shown in Table 1, the group of p_4 and p_2 works best. Compared to the single feature p_4+2conv+1pool, this fused feature brings in an average 1.65% performance improvement in *Recall-1*.

Table 2. The impact of weighting. FD denotes feature dimension. *Recall-1* is abbreviated as R1 and so forth. $*$ denotes the architecture of our DFCNN in Fig. 1.

Scale	FD	optical query			SAR query	
		R1	R5	mAP	R1	R5
(p_4+2conv)©(p_2+1conv+2pool)	675	0.690	0.805	0.647	0.681	0.798
(p_4+2conv)©(p_2+1conv+2pool+scale)	675	**0.694**	**0.815**	0.668	**0.685**	**0.809**
(p_4+2conv+1pool)©(p_2+1conv+3pool)	192	0.675	0.797	0.645	0.671	0.791
*** (p_4+2conv+1pool)© (p_2+1conv+3pool+scale)**	**192**	0.689	0.813	**0.671**	0.682	0.803

In our experiments, we add a scale layer to control the contribution of two types of features. As shown in Table 2, the weighting has a positive effect on all metrics. The lower-dimensional feature of 8×8 resolution (used by us) reaches comparable capacity with the feature of 15×15 resolution. The mAP rises by 2.6% and the *Recall* increases by an average of about 1.3%. Besides, the training methods are also important. When we change the training method of OED model from the dependence on semantic segmentation task to autoencoder model, the different pre-trained OptNet (*i.e.*, the encoder part of OED model) can be obtained. As shown in Table 3, compared to the use of pre-trained OptNet by autoencoder model, the result of fine-tuned Sub-network is better with

Table 3. The impact of different training methods. We always get the pre-trained SarNet by the autoencoder (AE) model. The pre-trained OptNet is obtained by semantic segmentation (SS) task or AE model.

Training Method	Feature Dimension	optical query				SAR query		
		R1	R5	R10	mAP	R1	R5	R10
Sub-network (SS + AE)	**115200**	**0.677**	**0.805**	**0.841**	**0.675**	**0.668**	**0.790**	**0.831**
Sub-network (AE + AE)	115200	0.382	0.580	0.651	0.379	0.392	0.594	0.662

the weights of OptNet trained by semantic segmentation task. It is because performing semantic segmentation task extracts more discriminative features.

Baseline and Proposed Method. In this paper, the effectiveness of SCCM algorithm [16] is demonstrated in two types of features. One is the 20000-D encoded SIFT features by bag-of-words model, the other is the CNN features learned by the DFCNN. And the result of CMCP algorithm [28] is based on the CNN features. Besides, the NetVLAD network [1] which works well for standard retrieval benchmarks, is taken as a branch of similar two-stream network. This network is initialized with the weights of Five-network in Fig. 3 and optimized with contrastive loss. Especially, we set 64 cluster centroids and using a trainable matrix reduces the feature dimensions into 1024 dimensions. According to [18], we implement the MISO method, which works on the identification of corresponding patches across optical and SAR images. The output of probability value is used to measure the similarity of two images here.

As shown in Table 4, the methods with deep features have obvious advantages over the methods with SIFT features. Besides, the result of SCCM method heavily relies on the basic features. Among three deep models, our method has the best performance. In addition, the output of DFCNN is 192-D features, whose dimension is lower than the output of NetVLAD. Moreover, the optical and SAR images can be processed separately in the DFCNN, but the MISO fails to do this. Hence, our method has the shortest inference time. The above analysis shows our retrieval method based on the DFCNN model is superior for handling this cross-modal task.

Table 4. The mAP of six methods when optical image as the query. The latter three methods are based on deep neural networks.

Method	SIFT	SCCM on		CMCP	**Ours**	MISO	NetVLAD
		SIFT	DFCNN				
mAP	0.004	0.029	0.669	0.393	**0.671**	0.600	0.262

4 Conclusion

In this paper, the cross-modal retrieval between optical and SAR remote sensing images is first proposed to our best knowledge. Considering the huge difference between optical and SAR images, we design an end-to-end deep model (*i.e.*, Double-Feature Convolutional Neural Network (DFCNN)) to solve this challenging issue. By adopting the tailored training strategy, the DFCNN can efficiently fuse two kinds of features after the channels-aggregated convolution and semi-average pooling (CS) operations. The CS scheme is a novel way to refine features. Quantitative experimental results demonstrate the effectiveness of the proposed cross-modal retrieval method. Furthermore, an optical/SAR image retrieval (OSR) dataset is collected for further researches. In the future, we intend to adjust the network model (*e.g.,* considering the multi-scale information) and extend our work to other research fields, such as image registration.

Acknowledgement. This work was supported in part by 973 Program under Contract 2015CB351803, by Natural Science Foundation of China (NSFC) under Contract 61390514 and 61331017, and by the Fundamental Research Funds for the Central Universities.

References

1. Arandjelovic, R., Gronat, P., Torii, A., Pajdla, T., Sivic, J.: Netvlad: CNN architecture for weakly supervised place recognition. In: CVPR, pp. 5297–5307 (2016)
2. Babenko, A., Lempitsky, V.: Aggregating local deep features for image retrieval. In: ICCV, pp. 1269–1277 (2015)
3. Badrinarayanan, V., Kendall, A., Cipolla, R.: Segnet: a deep convolutional encoder-decoder architecture for image segmentation. TPAMI **39**(12), 2481–2495 (2017)
4. Bell, S., Lawrence Zitnick, C., Bala, K., Girshick, R.: Inside-outside net: detecting objects in context with skip pooling and recurrent neural networks. In: CVPR, pp. 2874–2883 (2016)
5. Bottou, L.: Stochastic gradient descent tricks. In: Montavon, G., Orr, G.B., Müller, K.-R. (eds.) Neural Networks: Tricks of the Trade. LNCS, vol. 7700, pp. 421–436. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-35289-8_25
6. Cheng, G., Zhou, P., Han, J.: Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images. TGARS **54**(12), 7405–7415 (2016)
7. Chopra, S., Hadsell, R., LeCun, Y.: Learning a similarity metric discriminatively, with application to face verification. In: CVPR. vol. 1, pp. 539–546. IEEE (2005)
8. Eysenck, M.W., Keane, M.T.: Cognitive psychology: A student's handbook. Psychology press, New York (2013)
9. Fukui, K., Okuno, A., Shimodaira, H.: Image and tag retrieval by leveraging image-group links with multi-domain graph embedding. In: ICIP, pp. 221–225. IEEE (2016)
10. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: ICCV, pp. 1026–1034 (2015)
11. Hong, R., Hu, Z., Wang, R., Wang, M., Tao, D.: Multi-view object retrieval via multi-scale topic models. TIP **25**(12), 5814–5827 (2016)

12. Hong, R., Zhang, L., Tao, D.: Unified photo enhancement by discovering aesthetic communities from flickr. TIP **25**(3), 1124–1135 (2016)
13. Hong, R., Zhang, L., Zhang, C., Zimmermann, R.: Flickr circles: aesthetic tendency discovery by multi-view regularized topic modeling. TMM **18**(8), 1555–1567 (2016)
14. Jia, Y., et al.: Caffe: convolutional architecture for fast feature embedding. In: ACM MM, pp. 675–678. ACM (2014)
15. Li, Z., Li, Y., Gao, Y., Liu, Y.: Fast cross-scenario clothing retrieval based on indexing deep features. In: Chen, E., Gong, Y., Tie, Y. (eds.) PCM 2016. LNCS, vol. 9916, pp. 107–118. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-48890-5_11
16. Luo, M., Chang, X., Li, Z., Nie, L., Hauptmann, A.G., Zheng, Q.: Simple to complex cross-modal learning to rank. CVIU **163**, 67–77 (2017)
17. Mnih, V.: Machine learning for aerial image labeling. Ph.D. thesis, University of Toronto (Canada) (2013)
18. Mou, L., Schmitt, M., Wang, Y., Zhu, X.X.: A CNN for the identification of corresponding patches in SAR and optical imagery of urban scenes. In: JURSE, pp. 1–4. IEEE (2017)
19. Qi, Y., Song, Y.Z., Zhang, H., Liu, J.: Sketch-based image retrieval via siamese convolutional neural network. In: ICIP, pp. 2460–2464. IEEE (2016)
20. Rumelhart, D.E., Hinton, G.E., Williams, R.J.: Learning representations by back-propagating errors. Nature **323**(6088), 533 (1986)
21. Sharma, A., Jacobs, D.W.: Bypassing synthesis: PLS for face recognition with pose, low-resolution and sketch. In: CVPR, pp. 593–600. IEEE (2011)
22. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition (2014). arXiv preprint arXiv:1409.1556
23. Tambo, A.L., Bhanu, B.: Dynamic bi-modal fusion of images for the segmentation of pollen tubes in video. In: ICIP, pp. 148–152. IEEE (2015)
24. Wang, K., He, R., Wang, L., Wang, W., Tan, T.: Joint feature selection and subspace learning for cross-modal retrieval. TPAMI **38**(10), 2010–2023 (2016)
25. Wang, K., Yin, Q., Wang, W., Wu, S., Wang, L.: A comprehensive survey on cross-modal retrieval (2016). arXiv preprint arXiv:1607.06215
26. Wang, Y., Zhu, X.X., Zeisl, B., Pollefeys, M.: Fusing meter-resolution 4-d insar point clouds and optical images for semantic urban infrastructure monitoring. TGARS **55**(1), 14–26 (2017)
27. Wegner, J.D., Ziehn, J.R., Soergel, U.: Combining high-resolution optical and insar features for height estimation of buildings with flat roofs. TGARS **52**(9), 5840–5854 (2014)
28. Zhai, X., Peng, Y., Xiao, J.: Cross-modality correlation propagation for cross-media retrieval. In: ICASSP, pp. 2337–2340. IEEE (2012)
29. Zhong, P., Gong, Z., Li, S., Schönlieb, C.B.: Learning to diversify deep belief networks for hyperspectral image classification. TGARS **55**(6), 3516–3530 (2017)