

Weakly-Supervised Deep Image Hashing based on Cross-Modal Transformer

Ching-Ching Yang, Wei-Ta Chu
National Cheng Kung University, Tainan, Taiwan
chinx25425@gmail.com, wtchu@gs.ncku.edu.tw

Shiv Ram Dubey
Indian Institute of Information Technology, Allahabad, India
srdubey@iiita.ac.in

Abstract



Weakly-supervised image hashing emerges recently because web images associated with contextual text or tags are abundant. Text information weakly-related to images can be utilized to guide the learning of a deep hashing network. In this paper, we propose Weakly-supervised deep Hashing based on Cross-Modal Transformer (WHCMT). First, cross-scale attention between image patches is discovered to form more effective visual representations. A baseline transformer is also adopted to find self-attention of tags and form tag representations. Second, the cross-modal attention between images and tags is discovered by the proposed cross-modal transformer. Effective hash codes are then generated by embedding layers. WHCMT is tested on semantic image retrieval, and we show new state-of-the-art results can be obtained for the MIRFLICKR-25K dataset and NUS-WIDE dataset.

1 Introduction

Semantic image hashing provides a compact representation for efficiently measuring similarity between images. Most previous works determine the hash function in either a supervised or an unsupervised manner. In the supervised manner, ground truth labels associated with the images are used to measure semantic similarity. In the unsupervised manner, only images are considered, and the hash function is determined to minimize reconstruction errors or quantization errors. The supervised methods are limited by the small amount ground truth labels, while the unsupervised methods do not well utilize external information that can be easily obtained.

Table 1 shows three image samples and their associated tags and labels from the NUS-WIDE dataset [1]. Labels are usually from a predefined label set, and more accurately annotate images. On the other hand, tags are freely provided by users. The number of tags for an image is usually larger, and prone to different users' subjectivity. Though tags may be noisy, we may be able to utilize tag information to weakly supervise the learning process of a hashing function [2][3][4][5][6][7].

Table 1. Sample images and their associated tags and labels from the NUS-WIDE dataset.

Samples			
Labels	clouds, sky, structures	food	clouds, river, sky, water, sunset, lake, structures
Tags	#to1224, #j108, #london, #londres, #wesminster, #bigben, #geo:lat=51501093, #geo:lon=0124892, #geotagged	#chocolate, #cake, #chocolateganachebuttercream, #shamsd	#iceland, #icelandic, #reykjavik, #ice, #sky, #pink, #white, #blue, #clouds, #lake, #museum

A few issues arise: 1) Some tags may present weak semantics, but some do not. We need to model the correlation between tags and prioritize important ones in learning the hash space. 2) Not only tags, but also images may be noisy. Not all regions in an image present meaningful information. We thus need to find image regions that attend more to meaningful tags and utilize fused representations to learn a better hash space.

For weakly-supervised deep hashing, we introduce vision transformer [8] (ViT) and cross-modal attention to handle the aforementioned issues. ViT interprets images as sequences of patches and extracts visual representations. Fully-connected layers are then learnt to embed visual representations into hash codes [9]. It showed promising image retrieval performance based on hash codes learnt by ViT. But cross-modal attention in the weakly-supervised theme was not explored. Overall, contributions of this work include: 1) Weakly-supervised hashing: We propose to adopt ViT to do weakly-supervised hashing; 2) Cross-modal attention: We propose a cross-modal transformer to explore attention across visual representations and tag representations; 3) Evaluation: Our method outperforms the SOTA weakly-supervised hashing methods.

2 Cross-Modal Vision Transformer

2.1 Overview

Fig. 1 illustrates the proposed WHCMT. An image is viewed as a sequence of patches, like words in a sentence. A vision transformer can be developed to pro-

cess and embed image patches into visual representations. The meaning of a sentence can be represented in multiple ways based on different words. Similarly, we conceptually can represent the concept of an image in multiple ways based on patches of different scales. Inspired by CrossViT [10], we jointly consider patches at multiple scales (denoted as the L-branch and the S-branch in Fig. 1). For text information, tags are first embedded into vectors by the Word2Vec method [11]. A transformer is then developed to process tag vectors and then give rise to effective text representations.

In [5], visual representations and tag representations are jointly considered in loss functions that guide the learning of embedding layers for generating hash codes. In our work, we advocate that the cross-modal attention across image and tags should be discovered and well utilized. By taking visual rep. and tag rep. as “high-level tokens”, we further build a cross-modal transformer to find attention across modalities. The fused and attended representations are then fed to embedding layers to generate better hash codes.

2.2 Cross-Attention Vision Transformer

A vision transformer (ViT) projects image patches into tokens, denoted as \mathbf{x}_{patch} , by transformer encoders. A class token (\mathbf{x}_{cls}) is concatenated with the token sequence to interact with patch tokens, so that the processed class token conceptually summarizes information of the input image. To represent position information, ViT adds position embedding into each token, i.e., $\mathbf{x}_0 = [\mathbf{x}_{cls} || \mathbf{x}_{patch}] + \mathbf{x}_{pos}$, where the notation $||$ denotes vector concatenation. These tokens are then passed through a stack of transformer encoders to find self attention.

A transformer encoder is composed of a sequence of blocks. Each block consists of a multi-head self-attention (MSA) module and a feed-forward network (FFN). The process of the k th block is: $\mathbf{y}_k = \mathbf{x}_{k-1} + \text{MSA}(\text{LN}(\mathbf{x}_{k-1}))$ and $\mathbf{x}_k = \mathbf{y}_k + \text{FFN}(\text{LN}(\mathbf{y}_k))$, where LN denotes layer normalization.

In CrossViT [10], an image divided into patches of two granularities, and the class token from one granularity interacts with patch tokens at another granularity so that information at different levels is joined. As shown in Fig. 1, the larger (smaller) patch tokens \mathbf{x}_{patch}^l (\mathbf{x}_{patch}^s) are passed through the L-branch (S-branch).

Taking the L-branch as the main stream, the patch tokens from the S-branch are concatenated with the class token from the L-branch: $\mathbf{x}^{ll} = [f^l(\mathbf{x}_{cls}^l) || \mathbf{x}_{patch}^s]$, where $f^l(\cdot)$ is the projection function for dimension alignment. By taking only the class token \mathbf{x}_{cls}^l as the query, the transformer encoder discovers “cross-branch attention” between \mathbf{x}_{cls}^l and \mathbf{x}^{ll} , and is denoted as a cross attention (CA) module. Similar to MSA, multiple heads can be considered to form a multi-head cross

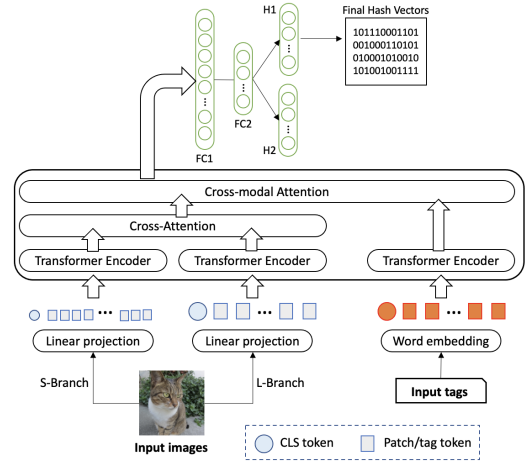


Figure 1. Framework of the proposed cross-modal transformer-based hashing.

attention MCA module.

$$\begin{aligned} \mathbf{y}_{cls}^l &= f^l(\mathbf{x}_{cls}^l) + \text{MCA}(\text{LN}([f^l(\mathbf{x}_{cls}^l) || \mathbf{x}_{patch}^s])), \\ \mathbf{v} &= [g^l(\mathbf{y}_{cls}^l) || \mathbf{x}_{patch}^l], \end{aligned} \quad (1)$$

where $g^l(\cdot)$ is also a projection function for dimension alignment. The output \mathbf{v} represents the final embedding of the input image after the multi-head cross attention process.

Specifically, we follow the settings proposed in CrossViT [10] to construct the L-branch and S-branch. The L-branch consists of a stack of three transformer encoders, while the S-branch is constituted by one transformer encoder. Therefore, precisely the input of the cross-attention module should be $\mathbf{x}^{ll} = [f^l(\mathbf{x}_{cls,3}^l) || \mathbf{x}_{patch,1}^s]$, where $\mathbf{x}_{cls,3}^l$ is the output of the class token given by the third transformer encoder in the L-branch, and $\mathbf{x}_{patch,1}^s$ is the output of the patch tokens given by the only transformer encoder in the S-branch. The description mentioned above is simplified for reading clarity.

2.3 Tag Transformer

Given a sequence of tags, a conventional transformer encoder is developed to obtain effective tag representations. A class token (\mathbf{t}_{cls}) is concatenated with the token sequence (\mathbf{t}_{tag}), i.e., $\mathbf{t}_0 = [\mathbf{t}_{cls} || \mathbf{t}_{tag}]$. Notice that tags are order-agnostic, and thus we don’t need positional embeddings. These tokens are then passed through a stack of transformer encoders, which components are the same as that mentioned in Eqn. (??). After processing, the output $\mathbf{t}_1 = [\mathbf{t}_{cls,1} || \mathbf{t}_{tag,1}]$ represents the embedded tag information after discovering multi-head self attention, where $\mathbf{t}_{cls,1}$ and $\mathbf{t}_{tag,1}$ denote the outputs corresponding to \mathbf{t}_{cls} and \mathbf{t}_{tag} , respectively. To simplify notation, the tag representation is denoted as \mathbf{t} in the following.

2.4 Cross-Modal Transformer

Inspired by cross attention inside a single modality, we find **cross-modal attention** between the image and tags. To discover cross-modal attention, we concatenate the visual representation $\mathbf{v} = [g^l(\mathbf{y}_{cls}^l) || \mathbf{x}_{patch}^l]$ with the tag representation $\mathbf{t} = [\mathbf{t}_{cls,1} || \mathbf{t}_{tag,1}]$, and then pass them to a transformer encoder with the multi-head cross-attention module: $\mathbf{a} = [\mathbf{v} || \mathbf{t}]$, $\mathbf{b} = \mathbf{a} + \text{MCA}(\text{LN}(\mathbf{a}))$, $\mathbf{a}' = \mathbf{b} + \text{FFN}(\text{LN}(\mathbf{b}))$.

The visual class token \mathbf{a}'_{cls} derived from $g^l(\mathbf{y}_{cls}^l)$ is concatenated with the image patch tokens \mathbf{x}_{patch}^l to form the final visual representation, i.e., $\mathbf{v}_f = [\mathbf{a}'_{cls} || \mathbf{x}_{patch}^l]$. Through discovering cross-modal attention, the patches and tags with higher correlations are prioritized to give more effective visual representations.

To generate hash codes, the visual representation \mathbf{v}_f is processed by two fully-connected (FC) layers (denoted as FC1 and FC2), and then passed to two layers called Head 1 (H1) and Head 2 (H2) in a lateral fashion, as illustrated in Fig. 1. The output of H1 is a b -dimensional hash code \mathbf{h} , which is the target result of the whole network and will be used in the downstream task like image retrieval. The output of H2 is a d -dim vector \mathbf{r} , which is compared with the processed tag tokens to force the network to form feature spaces in accordance with the semantic information brought by tags. This follows the design of WDHT [5].

2.5 Loss Functions

Primarily following [5], three losses are jointly considered to train the network: pairwise similarity loss, Hinge loss, and quantization loss. The pairwise similarity loss L_1 is designed to force that semantically similar images are mapped into similar hash codes. For any image pairs $(\mathbf{I}^{(i)}, \mathbf{I}^{(j)})$, L_1 is defined as:

$$L_1 = \sum_{i=1}^B \sum_{j=1}^B \left[\frac{1}{b} (\mathbf{h}^{(i)} - \mathbf{h}^{(j)})^T (\mathbf{h}^{(i)} - \mathbf{h}^{(j)}) - \frac{1}{2} \left(1.0 - \frac{\mathbf{t}_{cls}^{(i)T} \mathbf{t}_{cls}^{(j)}}{\|\mathbf{t}_{cls}^{(i)}\| \|\mathbf{t}_{cls}^{(j)}\|} \right) \right]^2, \quad (2)$$

where B is the mini-batch size, and $\mathbf{h}^{(i)}$ and $\mathbf{h}^{(j)}$ are the hash codes mapped from $\mathbf{I}^{(i)}$ and $\mathbf{I}^{(j)}$, respectively. The first term means the distance between hash codes of two images. It should be small if $\mathbf{I}^{(i)}$ is semantically similar to $\mathbf{I}^{(j)}$. The second term indicates the distance between two images' tag class tokens $\mathbf{t}_{cls}^{(i)}$ and $\mathbf{t}_{cls}^{(j)}$. If two images' tags are similar, the value $\frac{\mathbf{t}_{cls}^{(i)T} \mathbf{t}_{cls}^{(j)}}{\|\mathbf{t}_{cls}^{(i)}\| \|\mathbf{t}_{cls}^{(j)}\|}$ is larger, and the second term is smaller.

The output of H2 $\mathbf{r}^{(i)}$ for the i th image is forced to be similar to $\mathbf{t}_{cls}^{(i)}$. The similarity between $\mathbf{r}^{(i)}$ and $\mathbf{t}_{cls}^{(i)}$ should be larger than that between $\mathbf{r}^{(i)}$ and another image $\mathbf{I}^{(j)}$'s tags $\mathbf{t}_{cls}^{(j)}$ by the value *margin*. Specifically, L_2 in a mini-batch is defined as:

$$L_2 = \sum_i \sum_{j \neq i} \max[0, \text{margin} + \mathbf{t}_{cls}^{(j)T} \cdot \mathbf{r}^{(i)} - \mathbf{t}_{cls}^{(i)T} \cdot \mathbf{r}^{(i)}]. \quad (3)$$

The quantization loss L_3 is designed as:

$$L_3 = - \sum_{i=1}^B \frac{1}{b} (\mathbf{h}^{(i)} - 0.5\mathbf{1})^T \cdot (\mathbf{h}^{(i)} - 0.5\mathbf{1}), \quad (4)$$

where entries of the vector $\mathbf{1}$ is all ones. If the output of a neuron of H1 is closer to 0.5, it is penalized more.

Finally, the three losses are integrated as $L = \lambda_1 L_1 + \lambda_2 L_2 + \lambda_3 L_3$ to guide network training.

When training, the architecture shown in Fig. 1 jointly takes images (the left part) and corresponding tags (the right part) as the inputs to learn representations and hash codes. When doing retrieval, *only the test image is given and processed* (without tags), while the tag tokens are intentionally defined as vectors with all entries equal to one, i.e., 11111.... With this setting, the visual representations extracted by the left part are not affected by the right part after the cross-modal attention module, i.e., tag information is not used in retrieval for fair comparison.

3 Experiments

3.1 Evaluation Details

The NUS-WIDE dataset [1] and the MIR-FLICKR25K dataset [18] are used in the evaluation. We follow the evaluation protocol in [6], and please refer to it for detailed settings.

Regarding network training, as in [10], the cross attention module is pre-trained on the ImageNet1K dataset [19]. Based on the pre-trained CrossViT, we train from scratch for the rest of this framework, including the transformer encoders for tags and cross-modal attention, and the fully-connected layers to generate hash codes. The SGD algorithm is used to optimize the network parameters, with the learning rate as 0.001. The momentum rate is set to 0.9, and the size B of a mini-batch is set to 50. The weighting factors, λ_1 , λ_2 , and λ_3 are set to 1.0, 10.0, and 1.0, respectively. To get tag embeddings, the word2vec model pre-trained based on Wikipedia documents is adopted. A 300-dim embedding is generated for each tag.

The effectiveness of hash code learning is evaluated based on the task of semantic image retrieval. Mean average precision (MAP), precision, and recall are used to show performance. We compare with state-of-the-art unsupervised and weakly-supervised methods. Specifically, our proposed method is based on the framework similar to WDHT, and comparing with it demonstrates the effectiveness of cross-modal attention and CrossViT. WSDHQ is the most recent state of the art.

3.2 Performance Comparison

Table 2 shows MAP@5000 values of semantic image retrieval. Four observations can be made. First, weakly-supervised methods (from WMH to WHCMT)

Table 2. MAP values of semantic image retrieval, based on the top 5,000 retrieved images.

Method	MIR-FLICKR25K				NUS-WIDE			
	8 bits	16 bits	24 bits	32 bits	8 bits	16 bits	24 bits	32 bits
LSH [12]	0.524	0.570	0.562	0.572	0.376	0.392	0.413	0.418
SH [13]	0.592	0.609	0.617	0.604	0.498	0.505	0.477	0.492
SpH [14]	0.556	0.582	0.579	0.586	0.463	0.448	0.464	0.461
ITQ [15]	0.641	0.623	0.654	0.633	0.536	0.545	0.556	0.563
AQ [16]	0.637	0.645	0.658	0.661	0.524	0.567	0.587	0.592
DeepBit [17]	0.628	0.632	0.623	0.608	0.542	0.555	0.558	0.552
WMH [2]	0.656	0.684	0.672	0.671	0.558	0.592	0.605	0.601
WDH [3]	0.669	0.678	0.694	0.685	0.577	0.602	0.618	0.627
WDHT [5]	0.704	0.733	0.737	0.724	0.652	0.670	0.682	0.692
WSDHQ [6]	0.744	0.751	0.765	0.772	0.716	0.722	0.738	0.731
WHCMT	0.752	0.755	0.771	0.769	0.720	0.729	0.735	0.725

Table 3. Performance variations of different settings of weights to combine losses.

Settings	Weights	MAP@5000 (8 bits)
1	$\lambda_1 = 1, \lambda_2 = 10, \lambda_3 = 1$	0.755
2	$\lambda_1 = 1, \lambda_2 = 1, \lambda_3 = 1$	0.755
3	$\lambda_1 = 1, \lambda_2 = 0.1, \lambda_3 = 1$	0.753
4	$\lambda_1 = 1, \lambda_2 = 10, \lambda_3 = 0.1$	0.577
5	$\lambda_1 = 1, \lambda_2 = 10, \lambda_3 = 10$	0.752

outperform unsupervised methods (from LSH to DeepBit). This is not surprising because weakly-supervised methods obtain more clues (though noisy) from tags. Second, deep-based weakly-supervised methods (WDHT, WSDHQ, and our WHCMT) clearly obtain higher MAP values than non-deep methods (WMH and WDH). This shows higher capability of deep neural networks on modeling complex embedding. Third, our method significantly outperforms WDHT [5], which especially shows the effectiveness of cross-modal attention. Fourth, our method is competitive with the state-of-the-art WSDHQ [6]. Generally WHCMT performs better when fewer bits can be used to show hash codes.

3.3 Weightings for Losses

Table 3 shows performance variations of different weight settings to combine losses, based on 8-bits hash codes for the MIRFLICKR-25K dataset. By comparing setting #1 with settings #4 and #5, we see L_3 plays an important role because performance varies much if the corresponding weighting λ_3 is scaled 10 or 0.1 times. By comparing settings #1, #2, and #3, we see the loss L_2 relatively yields less influence.

3.4 Performance of Different Frameworks

To investigate the influence of different backbones to extract representations, we compare retrieval performance of several WDHT-based frameworks, as shown in Fig. 2. They are: (a) the original WDHT method with AlexNet to extract visual representations; (b) the WDHT method with the AlexNet replaced by ViT; (c) the WDHT method with the backbone replaced by CrossViT; and (d) the proposed WHCMT.

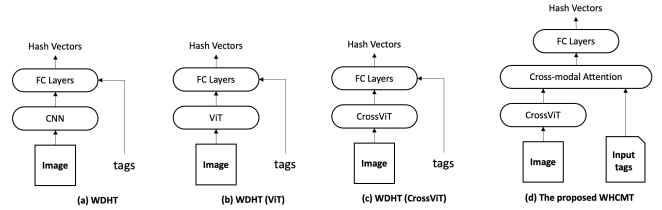


Figure 2. Illustrations of different comparison methods.

Table 4. Performance variations of different frameworks.

Method	MAP@5000 (8 bits)	Improvement (%)
WDHT (original) [5]	0.704	—
WDHT (ViT)	0.729	3.55%
WDHT (CrossViT)	0.733	4.12%
WHCMT	0.755	7.24%

Table 4 shows performance variations of different approaches. By replacing AlexNet with ViT or CrossViT, performance is boosted. By further considering cross-modal attention, our WHCMT outperforms WDHT by a significant margin, which verifies the value of cross-modal attention.

4 Conclusion

We discover cross-modal attention and cross-scale attention for weakly-supervised image hashing. Information of image patches at different scales can be cross referred based on a vision transformer. In addition, attention across modalities can further enhance representations. We thus propose a cross-modal transformer to form more effective representations, which are then used to generate better hash codes. Experimental results show that the proposed WHCMT method significantly outperforms the SOTAs on two major datasets.

Acknowledgement. This work was funded in part by Qualcomm through a Taiwan University Research Collaboration Project and in part by the National Science and Technology Council, Taiwan, under grants 112-2425-H-006-001, 111-3114-8-006-002, 111-2634-F-006-012, and 110-2221-E-006-127-MY3.

References

- [1] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng, “Nus-wide: A real-world web image database from national university of singapore,” in *Proceedings of ACM International Conference on Image and Video Retrieval*, 2009.
- [2] Jinhui Tang and Zechao Li, “Weakly supervised multimodal hashing for scalable social image retrieval,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 10, pp. 2730–2741, 2018.
- [3] Hui Cui, Lei Zhu, Chaoran Cui, Xiushan Nie, and Huaxiang Zhang, “Efficient weakly-supervised discrete hashing for large-scale social image retrieval,” *Pattern Recognition Letters*, vol. 130, pp. 174–181, 2020.
- [4] Lu Jin, Zechao Li, Yonghua Pan, and Jinhui Tang, “Weakly-supervised image hashing through masked visual-semantic graph-based reasoning,” in *Proceedings of ACM International Conference on Multimedia*, 2020, pp. 916–924.
- [5] Vijetha Gattupalli, Yaoxin Zhuo, and Baoxin Li, “Weakly supervised deep image hashing through tag embeddings,” in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2019.
- [6] Jinpeng Wang, Bin Chen, Qiang Zhang, Zaiqiao Meng, Shangsong Liang, and Shutao Xia, “Weakly supervised deep hyperspherical quantization for image retrieval,” in *Proceedings of AAAI Conference on Artificial Intelligence*, 2021, pp. 2755–2763.
- [7] Lei Zhu, Hui Cui, Zhiyong Cheng, Jingjing Li, and Zheng Zhang, “Dual-level semantic transfer deep hashing for efficient social image retrieval,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 4, pp. 1478–1489, 2021.
- [8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Szekoreit, and Neil Houlsby, “An image is worth 16x16 words: Transformers for image recognition at scale,” in *Proceedings of International Conference on Learning Representations*, 2021.
- [9] Shiv Ram Dubey, Satish Kumar Singh, and Wei-Ta Chu, “Vision transformer hashing for image retrieval,” in *arXiv:2109.12564*, 2021.
- [10] Chun-Fu Chen, Quanfu Fan, and Rameswar Panda, “Crossvit: Cross-attention multi-scale vision transformer for image classification,” in *Proceedings of IEEE International Conference on Computer Vision*, 2021, pp. 357–366.
- [11] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean, “Efficient estimation of word representations in vector space,” in *Proceedings of International Conference on Learning Representations*, 2013.
- [12] Moses S. Charikar, “Similarity estimation techniques from rounding algorithms,” in *Proceedings of ACM symposium on Theory of Computing*, 2002, pp. 380–388.
- [13] Yair Weiss, Antonio Torralba, and Rob Fergus, “Spectral hashing,” in *Proceedings of Advances in Neural Information Processing Systems*, 2009, pp. 1753–1760.
- [14] Jae-Pil Heo, Youngwoon Lee, Junfeng He, Shih-Fu Chang, and Sung-Eui Yoon, “Spherical hashing,” in *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, 2012, pp. 2957–2964.
- [15] Yunchao Gong, Svetlana Lazebnik, Albert Gordo, and Florent Perronnin, “Iterative quantization: A procrustean approach to learning binary codes for large-scale image retrieval,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 12, pp. 2916–2929, 2013.
- [16] Artem Babenko and Victor Lempitsky, “Additive quantization for extreme vector compression,” in *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, 2014, pp. 931–938.
- [17] Kevin Lin, Jiwen Lu, Chu-Song Chen, and Jie Zhou, “Learning compact binary descriptors with unsupervised deep neural networks,” in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1183–1192.
- [18] Mark J. Huiskes and Michael S. Lew, “The mir flickr retrieval evaluation,” in *Proceedings of ACM International Conference on Multimedia Retrieval*, 2008, pp. 39–43.
- [19] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.