# USING DEEP CROSS MODAL HASHING AND ERROR CORRECTING CODES FOR IMPROVING THE EFFICIENCY OF ATTRIBUTE GUIDED FACIAL IMAGE RETRIEVAL

*Veeru Talreja*[*]*, Fariborz Taherkhani*[*]*, Matthew C. Valenti , and Nasser M. Nasrabadi*

West Virginia University, Morgantown, USA

## ABSTRACT

With benefits of fast query speed and low storage cost, hashing-based image retrieval approaches have garnered considerable attention from the research community. In this paper, we propose a novel Error-Corrected Deep Cross Modal Hashing (CMH-ECC) method which uses a bitmap specifying the presence of certain facial attributes as an input query to retrieve relevant face images from the database. In this architecture, we generate compact hash codes using an end-to-end deep learning module, which effectively captures the inherent relationships between the face and attribute modality. We also integrate our deep learning module with forward error correction codes to further reduce the distance between different modalities of the same subject. Specifically, the properties of deep hashing and forward error correction codes are exploited to design a cross modal hashing framework with high retrieval performance. Experimental results using two standard datasets with facial attributes-image modalities indicate that our CMH-ECC face image retrieval model outperforms most of the current attribute-based face image retrieval approaches.

***Index Terms*—** Cross-modal hashing, deep learning, facial attributes, error correcting codes, standard array

## 1. INTRODUCTION

With the fast development of search engines and social networks, there exists a vast amount of multimedia data, such as texts, images and videos being generated on the world wide web everyday. The presence of multimedia big data has sparked a rise of content based image retrieval (CBIR) techniques in the research community. Approximate nearest neighbors (ANN) based semantic search has garnered a lot of attention to guarantee the retrieval quality and computing efficiency for CBIR in large-scale datasets. Cross-modal retrieval is an important paradigm of CBIR, which works with multimodal data and supports similarity retrieval across different modalities, e.g., retrieval of relevant facial images in response to attribute query such as "an old woman wearing glasses". In this paper, we address the problem of cross-modal retrieval of relevant face images in response to facial attributes queries by utilizing a deep cross-modal hashing framework in combination with error correcting codes.

A fast and promising solution to ANN search for cross-modal retrieval is cross-modal hashing (CMH), which compresses high-dimensional data into compact binary codes and maintains the semantic similarity by mapping images of similar content to similar binary codes. CMH returns relevant results of one modality in response to query of another modality, where respective hash codes in the same latent Hamming space are generated for each individual modality. Recently, application of deep learning to hash methods for uni-modal image retrieval [1, 2] and cross-modal retrieval [3, 4] have shown that end-to-end learning of feature extraction and hash coding using deep neural networks is more efficient than using the hand-crafted features [5, 6]. Particularly, it proves beneficial to jointly learn semantic similarity preserving features and also curb the quantization error of binarizing continuous representation to hash codes.

Searching for facial images of people including identification in response to a facial attribute query has been investigated in the past [7, 8, 9, 10]. However, all of these methods use hand-crafted features to perform a cross-modal retrieval. We present a novel CMH framework called CMH-ECC for error-corrected attribute guided deep cross-modal hashing for face-image retrieval from large datasets. The main contributions of this paper include: (1) **Error-corrected attribute guided deep cross modal hashing (hereon known as CMH-ECC)** : We have designed a novel architecture using deep cross modal hashing for face image retrieval in response to an attribute query. (2) **Error correcting codes**: We have integrated the deep cross modal hashing with error correcting codes to further reduce the Hamming distance between different modalities of same subject and improve the retrieval efficiency obtained from performing only deep cross modal hashing. (3) **Scalable cross-modal hash**: Our architecture CMH-ECC performs facial image retrieval using point wise data without requiring pairs or triplets of training inputs, which makes CMH-ECC scalable to large scale datasets.

[*] Authors Contributed Equally

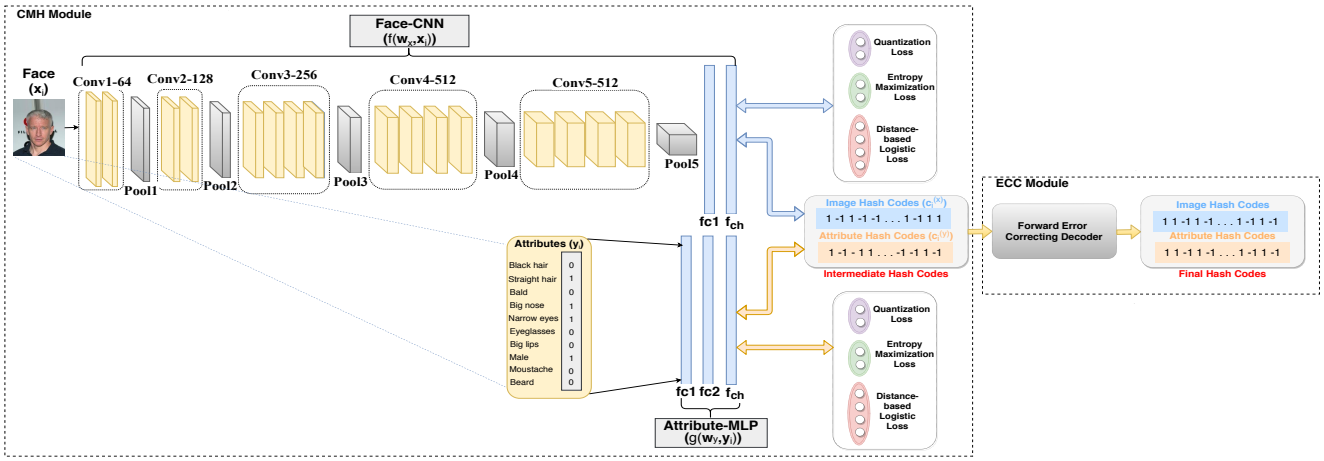



**Fig. 1**: Block Diagram of the CMH-ECC.

## 2. THE PROPOSED CMH-ECC FRAMEWORK

### 2.1. Problem Definition

Define $\mathcal{O} = \{\mathbf{o}_i\}_{i=1}^{n}$ to be the training set where $n$ is the number of training samples. All the samples have two modalities $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^{n}$ and $\mathbf{Y} = \{\mathbf{y}_i\}_{i=1}^{n}$, which corresponds to image and attribute modalities, respectively. $\mathbf{x}_i$ is the raw image $i$ in a training set of size $n$ and $\mathbf{y}_i$ is the annotated facial attributes vector related to image $i$. $\mathbf{S}$ is a cross-modal similarity matrix in which $S_{ij} = 1$ if image $\mathbf{x}_i$ contains a $y_j$ facial attribute, and $S_{ij} = 0$ otherwise.

Based on the given training information (i.e., $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{S}$), the proposed method learns two modality-specific hashing functions: $h^{(x)}(\mathbf{x}) \in \{-1, +1\}^d$ for image modality and $h^{(y)}(\mathbf{y}) \in \{-1, +1\}^d$ for attribute modality where $d$ is the number of the bits used in the intermediate hash codes. The two hashing functions have to preserve the cross-modal similarity in $\mathbf{S}$. Specifically, if $S_{ij} = 1$, the Hamming distance between the binary codes $\mathbf{c}_i^{(x)} = h^{(x)}(\mathbf{x}_i)$ and $\mathbf{c}_j^{(y)} = h^{(y)}(\mathbf{y}_j)$ should be small and if $S_{ij} = 0$, the corresponding Hamming distance should be large. The learned hash functions can be employed to generate $d$-bit intermediate hash codes for query and database instances in both modalities. The intermediate hash codes for query and database points are passed through a forward error correcting (FEC) decoder $f^{(d)}(.)$ to generate the final $c$-bit codewords (final hash codes) which are used in the retrieval process.

The block diagram of the proposed framework is given in Fig. 1. The proposed CMH-ECC framework has two modules. The first module is the deep cross modal hashing module (CMH module) and the second module in the CMH-ECC is the error correcting code module (ECC module).

### 2.2. Deep cross-modal hashing module (CMH)

CMH module trains a coupled deep neural network (DNN) to generate intermediate hash codes using a distance-based logistic loss to preserve the cross-modal similarity. The CMH module has three main functions: 1) Learn a coupled DNN using distance-based logistic loss to preserve the cross-modal similarity. 2) In order to preserve a high retrieval performance, control the quantization error for each modality due to the binarization of continuous output activations of the network to hash codes. 3) Maximize the entropy corresponding to each bit to obtain the maximum information provided by the hash codes.

The CMH module is composed of two networks : A Convolutional Neural Network (CNN) to extract features for image modality and a Multi-Layer Perceptron (MLP) to extract features for facial attribute modality. For CNN network, we have used VGG-19 [11] network pre-trained on the ImageNet [12] dataset as a starting point and fine-tuned it as a classifier by using the CASIA-Web Face dataset. The original VGG-19 consists of five convolutional layers ($conv1 - conv5$) and three fully-connected layers ($fc6 - fc8$). We discard the $fc8$ layer and replace the $fc7$ layer with a new $f_{ch}$ layer with $d$ hidden nodes, where d is the required intermediate hash code length (the intermediate code length in all the experiments is set to 256 bits). The MLP network comprises three fully connected layers to represent features for the facial attribute modality. To learn attribute features from this network, we annotate each training sample image with a binary bit map that indicates the presence or absence of corresponding facial attribute. This bitmap serves as a facial attribute vector and is used as input to the MLP network. The first and second layers in the MLP network contain 4,096 nodes with ReLU activation and the number of nodes in the last fully connected layer is equal to the intermediate hash code length $d$ with identity activation. We use the Adam optimizer [13] with the default hyper-parameter values ($\epsilon = 10^{-3}$, $\beta_1 = 0.9$, $\beta_2 = 0.999$) to train all the parameters using alternative minimization approach. The batch size in all the experiments is fixed to 128.

For efficient retrieval results, assuming that two samples $o_i$ and $o_j$ are semantically similar, their corresponding hash


**Fig. 2**: Qualitative results: Retrieved images using CMH-ECC for given facial attributes.

codes should also be similar in the low dimensional Hamming space. We design the objective function for generating efficient hash codes. Our objective function for CMH comprises of three parts: (1) distance-based logistic loss; (2) quantization loss; and (3) entropy maximization loss.

Let $f(\mathbf{w}_x, \mathbf{x}_\text{i}) \in \mathbb{R}^d$ and $g(\mathbf{w}_y, \mathbf{y}_\text{j})$ represent the learned CNN features for image modality $\mathbf{x}_\text{i}$ and MLP features for attribute modality $\mathbf{y}_\text{j}$, respectively. $\mathbf{w}_x$ and $\mathbf{w}_y$ are the CNN network weights and the MLP network weights, respectively. We define the total objective function for CMH as follows:

$$\min_{\mathbf{C}_{x,y},\mathbf{w}_x,\mathbf{w}_y} \mathcal{J} = \sum_{i=1}^{n}\sum_{j=1}^{n} \underbrace{\ell_c(p(\mathbf{F}_{*i}, \mathbf{G}_{*j}), S_{ij})}_{\text{distance- based logistic loss}} + \underbrace{\alpha\,(||\mathbf{F} - \mathbf{C}_x||_F^2 + ||\mathbf{G} - \mathbf{C}_y||_F^2)}_{\text{quantization loss}} + \underbrace{\beta\,(||\mathbf{F1}||_F^2 + ||\mathbf{G1}||_F^2)}_{\text{entropy maximization}} \quad \text{s.t.} \quad \mathbf{C}_{x,y} \in \{+1, -1\}^{d\times n}, \tag{1}$$

where $\mathbf{F} \in \mathbb{R}^{d\times n}$ is the image feature matrix and $\mathbf{G} \in \mathbb{R}^{d\times n}$ is the facial attribute feature matrix constructed by placing column-wise CNN and MLP features of training samples respectively. $\mathbf{F}_{*i} = f(\mathbf{w}_x, \mathbf{x}_\text{i})$ is the CNN feature corresponding to sample $\mathbf{x}_\text{i}$ and $\mathbf{G}_{*j} = f(\mathbf{w}_y, \mathbf{y}_\text{j})$ is the MLP feature corresponding to sample $\mathbf{y}_\text{j}$. $\mathbf{C}_x$ and $\mathbf{C}_y$ are the binary hash code matrices for image and attribute modalities, respectively. Notation $\mathbf{1}$ represents a vector with all its elements set to 1.

The first term in the objective function is distance-based logistic loss. This loss causes modalities referring to the same sample to attract one another and modalities to repel if they refer to two different samples. The distance based logistic-loss is derived from distance-based logistic probability, which is given by $p(\mathbf{F}_{*i}, \mathbf{G}_{*j}) = \frac{1+exp(-m)}{1+exp(||\mathbf{F}_{*i}-\mathbf{G}_{*j}||-m)}$ and represents the probability of the match between the image modality feature vector $\mathbf{F}_{*i}$ and attribute modality feature vector $\mathbf{G}_{*j}$, given their squared distance. The margin parameter $m$ determines the extent to which matched or non-matched samples are attracted or repelled, respectively. Then we apply the cross entropy loss similar to the classification case for deriving the final distance-based logistic loss : $\ell_c(p, s) = -slog(p) + (s-1)log(1-p)$. The second term in the objective function helps us to preserve the cross-modal similarity in the binary domain using hash codes $\mathbf{C}_x$ and $\mathbf{C}_y$, where $\mathbf{C}_x = \text{sign}(\mathbf{F})$ and $\mathbf{C}_y = \text{sign}(\mathbf{G})$. The third term in the objective function attempts to maximize the entropy on the bits of the hash code by making each bit of the hash code be balanced on all the training points. Precisely, the number of +1 and 1 for each bit on all the training samples should be almost the same. $\alpha$ and $\beta$ are tuning parameters that we set to 1.

### 2.3. Error correcting code module (ECC)

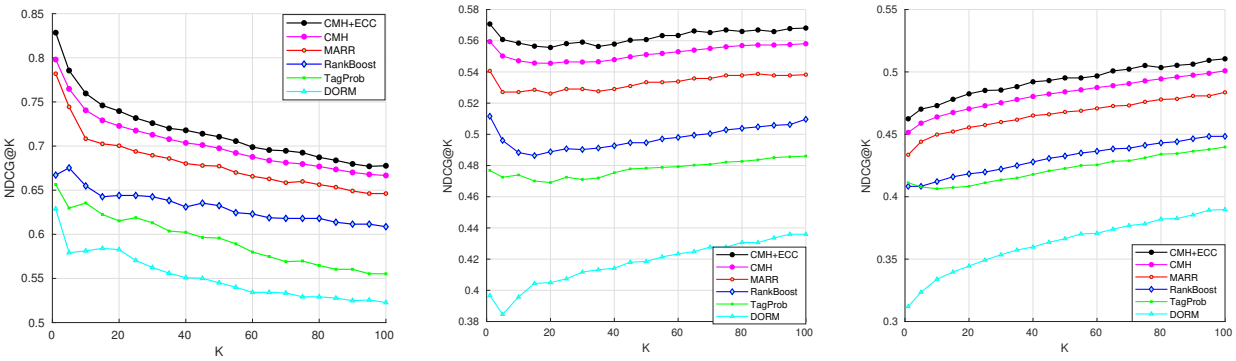
The intermediate hash codes generated by the CMH module can be used for a retrieval process. However, after gaining experience with CMH, we have concluded that there is an opportunity for improvement and further reducing the Hamming distance for different modalities of the same subject. On further research and inspired by [14], we identified error correcting codes to be a promising solution for reducing the Hamming distance for different modalities of the same subject.

We assume that the intermediate hash code generated by the CMH module is a binary vector that is within a certain distance from a codeword of an error-correcting code. By passing the intermediate hash code through an appropriate FEC decoder, the closest codeword is found and this closest codeword is used as a final hash code for the retrieval process. The main component of the ECC module is the forward error correcting (FEC) decoder. Due to their minimum-distance separable (MDS) property and widely available hardware, we have adopted Reed Solomon (RS) codes as our form of coding for FEC decoder.

The RS codes use symbols of length $m$ bits. Using a symbol size of $m$ bits, the length of the RS codeword is given by $N = 2^{m-1}$ in symbols, which corresponds to $n = mN$ in bits. However, we have utilized shortened RS codes for designing the FEC decoder of the ECC module. A shortened RS code is one in which the input to the decoder given as $N_1$ is less than the actual codeword length $N = 2^{m-1}$. For our decoder, we have used shortened RS code with $m = 8$ and $N = 255$ symbols and the input to the decoder $N_1$ equal to 32 symbols which is equal to 256 bits. The intermediate hash code, which is used as the input to the FEC decoder $N_1$ is taken as 256 bits for all the experiments. The codewords generated after decoding correspond to the final hash codes which are used for retrieval process using Hamming distance.

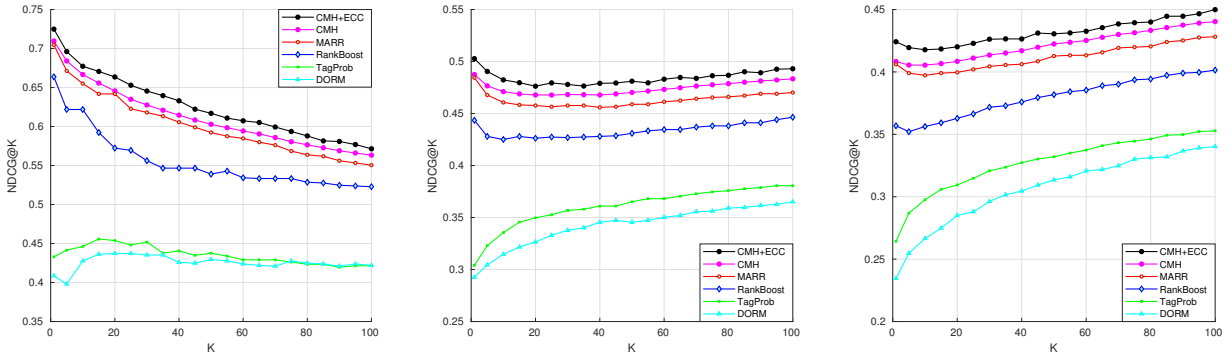
## 3. EXPERIMENTAL RESULTS

**Datasets**: FaceTracer [7] and LFW [15] datasets have been used to evaluate our proposed framework. LFW is a popular dataset of more than 13,000 images of faces collected from




(a) Single Attribute Queries (b) Double Attribute Queries (c) Triple Attribute Queries

**Fig. 3**: Ranking performance on the LFW dataset.




(a) Single Attribute Queries (b) Double Attribute Queries (c) Triple Attribute Queries

**Fig. 4**: Ranking performance on the FaceTracer dataset.

the internet for face recognition as well as attribute classification. The FaceTracer dataset is a large collection of $15,000$ real-world face images, collected from the internet.

**Evaluation Results**: We follow the experimental protocol used in Multi Attribute Retrieval and Ranking (MARR) [9]. We use normalized discounted cumulative gain (NDCG) as our evaluation metric to compare CMH-ECC performance with other methods. NDCG is a standard single-number measure of ranking quality that allows non-binary relevance judgments, while most traditional ranking measures only allow binary relevance (relevant or not relevant). NDCG is defined as $NDCG@k = \frac{1}{Z}\sum_{i=1}^{k}\frac{2^{rel(i)}-1}{log(i+1)}$, where rel(i) is the relevance of the $i^{th}$ ranked image and Z is a normalization constant to ensure that the correct ranking results in an NDCG score of 1.

Fig. 2 indicates the qualitative result of CMH-ECC approach for the given facial attributes. We compare the ranking quality using NDCG scores of the proposed CMH-ECC with four state of the art retrieval methods including MARR [9], rankBoost [16], Direct Optimization of Ranking Measures (DORM) [17], TagProp [18]. In addition, we have also compared our results of CMH-ECC framework with the results of using only CMH module without FEC, which implies using intermediate hash codes as our final hash codes. Fig. 3 and Fig. 4 plots the NDCG scores, as a function of the ranking truncation level $k$, using different number of attribute queries for the LFW and FaceTracer dataset, respectively. We can observe that CMH-ECC generally outperforms the comparison methods for both datasets using all the three types of queries. In particular, compared to the state of the art method MARR, we achieve approximately an increase of $4.0\%$, $3.5\%$ and $3.0\%$ in NDCG values for single, double and triple attribute queries, respectively. The retrieval efficiency using intermediate hash codes generated by our CMH module also outperforms MARR. Notice that NDCG values for the FaceTracer dataset for all the methods are relatively lower when compared to the LFW dataset. This is due to the difference in the distributions of the two datasets.

## 4. CONCLUSION

In this paper, we proposed a facial retrieval algorithm using deep hashing network and forward error correcting decoder to retrieve relevant facial images from the database using a given attribute query. This is the first time where error correcting codes have been combined with deep cross modal hashing for image retrieval. The experimental results on two popular public datasets show that our method outperforms the current face image retrieval approaches in the literature.

## 5. REFERENCES


[1] H. Liu, R. Wang, S. Shan, and X. Chen, "Deep supervised hashing for fast image retrieval," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016, pp. 2064–2072.

[2] Zhangjie Cao, Mingsheng Long, Jianmin Wang, and Philip S Yu, "Hashnet: Deep learning to hash by continuation," *arXiv preprint arXiv:1702.00758*, 2017.

[3] Erkun Yang, Cheng Deng, Wei Liu, Xianglong Liu, Dacheng Tao, and Xinbo Gao, "Pairwise relationship guided deep hashing for cross-modal retrieval," in *AAAI*, 2017.

[4] Qing-Yuan Jiang and Wu-Jun Li, "Deep cross-modal hashing," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3232–3240.

[5] Dongqing Zhang and Wu-Jun Li, "Large-scale supervised multimodal hashing with semantic correlation maximization," in *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*. 2014, AAAI'14, pp. 2177–2183, AAAI Press.

[6] Yi Zhen and Dit-Yan Yeung, "Co-regularized hashing for multimodal data," in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds., pp. 1376–1384. Curran Associates, Inc., 2012.

[7] Neeraj Kumar, Peter Belhumeur, and Shree Nayar, "Facetracer: A search engine for large collections of images with faces," in *European conference on computer vision*. Springer, 2008, pp. 340–353.

[8] D. A. Vaquero, R. S. Feris, D. Tran, L. Brown, A. Hampapur, and M. Turk, "Attribute-based people search in surveillance environments," in *2009 Workshop on Applications of Computer Vision (WACV)*, Dec 2009, pp. 1–8.

[9] B. Siddiquie, R. S. Feris, and L. S. Davis, "Image ranking and retrieval based on multi-attribute queries," in *CVPR 2011*, June 2011, pp. 801–808.

[10] Fariborz Taherkhani, Nasser M Nasrabadi, and Jeremy Dawson, "A deep face identification network enhanced by facial attributes prediction," *2018 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2018.

[11] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2009*. IEEE, 2009, pp. 248–255.

[13] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[14] V. Talreja, M. C. Valenti, and N. M. Nasrabadi, "Multibiometric secure system based on deep learning," in *2017 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, Nov 2017, pp. 298–302.

[15] Gary B Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," Tech. Rep., Technical Report 07-49, University of Massachusetts, Amherst, 2007.

[16] Yoav Freund, Raj Iyer, Robert E Schapire, and Yoram Singer, "An efficient boosting algorithm for combining preferences," *Journal of machine learning research*, vol. 4, no. Nov, pp. 933–969, 2003.

[17] Quoc V. Le and Alexander J. Smola, "Direct optimization of ranking measures," *CoRR*, vol. abs/0704.3359, 2007.

[18] M. Guillaumin, T. Mensink, J. Verbeek, and C. Schmid, "Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation," in *2009 IEEE 12th International Conference on Computer Vision*, Sept 2009, pp. 309–316.