Weakly Supervised Image Retrieval via Coarse-scale Feature Fusion and Multi-level Attention Blocks

Xinyao Nie, Hong Lu, Zijian Wang, Jingyuan Liu, Zehua Guo
{xynie18,honglu,zijianwang18,jingyuanliu15,zehuaguo16}@fudan.edu.cn
Shanghai Key Laboratory of Intel. Info. Processing, School of Computer Science
Fudan University, P. R. China

ABSTRACT

In this paper, we propose an end-to-end Attention-Block network for image retrieval (ABIR), which greatly increases the retrieval accuracy without human annotations like bounding boxes. Specifically, our network utilizes coarse-scale feature fusion, which generates the attentive local features via combining the information from different intermediate layers. Detailed feature information is extracted with the application of two attention blocks. Extensive experiments show that our method outperforms the state-of-the-art by a significant margin on four public datasets for image retrieval tasks.

KEYWORDS

Image retrieval, weakly supervised, attention block, feature fusion

ACM Reference Format:
Xinyao Nie, Hong Lu, Zijian Wang, Jingyuan Liu, Zehua Guo. 2019. Weakly Supervised Image Retrieval via Coarse-scale Feature Fusion and Multi-level Attention Blocks. In *2019 International Conference on Multimedia Retrieval (ICMR'19), June 10–13, 2019, Ottawa, ON, Canada.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3323873.3325017

1 INTRODUCTION

Recently, CBIR [26] has gained great popularity in both academic community and industry, since image search (especially image-by-image search service) is needed in almost every search engine and it can be directly applied to various applications such as product recognition [5, 8, 24], webscale image retrieval, and face recognition [14]. In this paper, we focus on the task of improving the accuracy of CBIR. Specifically, since networks trained by large amount of annotated data commonly suffer from the difficulty of generalization and the process of building large-scale annotated datasets is both costly and prone to errors, our method makes use of weakly supervised mechanisms like [2, 21, 23], which acquire results without additional annotations like bounding boxes, landmarks, attributes, etc.

With the help of Convolutional Neural Network (CNN), which is sufficiently discriminative without any embedding and complex aggregation in manually crafted feature, researchers have made significant improvements on this task, such as ImageNet [16], FashionNet [10], etc. As accurate CBIR requires reasonable and effective image feature extraction approaches, a hybrid method which extracts multiple CNN features from the input image was presented by Razavian et al. [11]. However, the efficiency in feature computation and encoding steps was also indispensable, which results in the application of pre-trained CNN models. Ng et al. [9] and Tolias et al. [18] used the column features based on the pre-trained CNN model for instance. Although pre-trained CNN models have achieved impressive retrieval accuracy, a popular topic consists in fine-tuning the CNN model on a specific dataset. Therefore, Babenko et al. first fine-tuned a CNN model for generic instance image retrieval [1]. The aforementioned approaches have shown strong performances on feature extraction and representation for CBIR. Nevertheless, they always suffer from the limitation of further improving accuracy of image retrieval because it is difficult to generate expressive and general representation with multi-dimension features in objects like cars or birds. Another topic interests us is fashion retrieval. It is a hard one because detecting and identifying clothing in an unconstrained image relies on fine-grained information, which could be a problem even for humans. Thus, we need to introduce a mechanism to let the network focus on the related regions and also combine features more efficiently.

To address the mentioned problems, we propose an end-to-end Attention-Block network for Image Retrieval (ABIR), which is mainly inspired by [12] and [13]. [12] has introduced a multi-level attention mechanism for object classification, showing that the proposed attention maps can be applied to weakly-supervised object localization. [13] has proposed a grid attention mechanism aiming to explore more detailed information from the intermediate layers. To be specific, ABIR is designed based on the popular CNN architecture of VGG16 [15], combining the low-level features into abstract high-level features to learn the semantic representation. To make the model eligible in different scenes and amplify the useful information in images, we also propose two novel attention blocks, which are used to generate the attentive local features via combining the information from VGG16's different intermediate layers. One of the attention block focuses on the surroundings of the query image while the other on object parts. Comprehensive evaluations on four large-scale datasets [7, 8, 17, 20] demonstrate the superiority of our proposed method.

In summary, our main contributions are as follows:

1. We propose an ABIR: the end-to-end Attention-Block network that addressed image retrieval with weakly supervised learning and less number of parameters, via fusing multi-dimension features and incorporating a concise but helpful loss function.

ICMR'19, June 10–13, 2019, Ottawa, ON, Canada
© 2019 Association for Computing Machinery.
ACM ISBN 978-1-4503-6765-3/19/06…$15.00
https://doi.org/10.1145/3323873.3325017

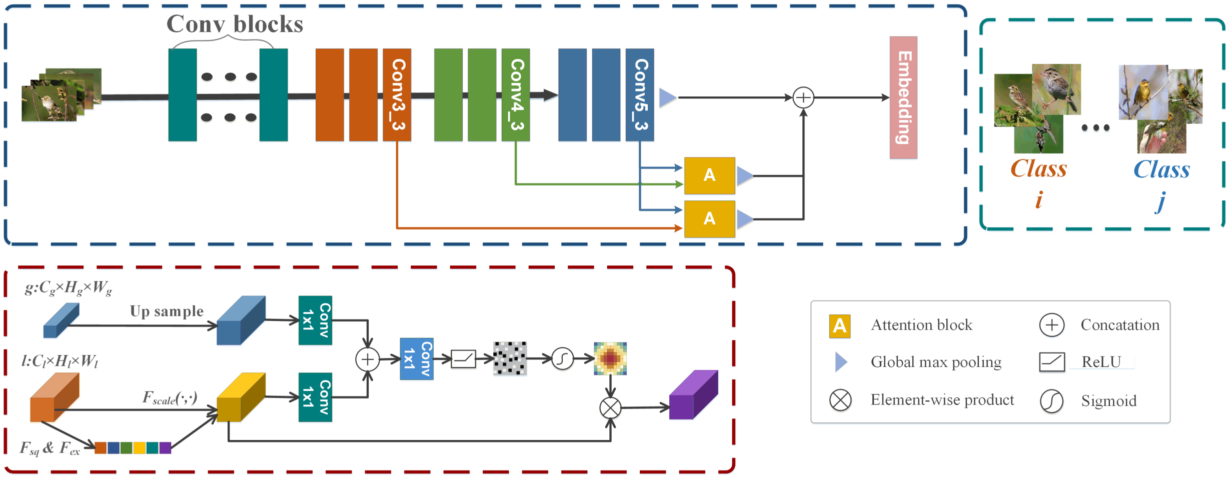

Figure 1: Overview of our network architecture and attention block. The first attention block combines the information from Conv3_3 and Conv5_3 layers, while the second attention block combines the information from Conv4_3 and Conv5_3 layers.

2. We show the benefit of training with the two attention blocks, which highlight interest areas on the input image and is powerful to automatic information filter.

3. Quantitatively, we report large improvement over the state-of-the-art on image retrieval across four public datasets [7, 8, 17, 20].

2 PROPOSED METHOD

In this section, we present our proposed method in detail. We give a brief description of our network and its whole architecture at first. Then we introduce the key part of our ABIR, multi-level attention module, which could automatically and precisely attend to discriminative parts of target objects during weakly-supervised training. Consequently, the network is able to learn more distinctive information from the original images for efficient feature selection.

2.1 Network Architecture

Attention-Block network for Image Retrieval (ABIR) takes an image as input and extracts feature maps of conv5_3 layers as global features. Next, the feature maps of conv3_3 and conv4_3 are combined with conv5_3 into two attention blocks, from which the attentive local features are generated. Then the network is designed to produce the global and local feature vectors by employing global max pooling operation. Finally, a concatenation of these feature vectors is fed to embedding layer at the end of the pipeline and thus producing the representation vector of this image. Our ABIR architecture is presented in Figure 1.

2.2 Attention Block

Following the variable representation of [12], $\mathcal{L}^s = \left\{\boldsymbol{l}_i^s\right\}_{i=1}^n$ are the feature vectors extracted from layer $s \in \{1, ..., S\}$. Each $\boldsymbol{l}_i^s$ denotes the feature vector of the i th spatial map in layer s. Let $\boldsymbol{g} \in \mathbb{R}^{C_g}$ (C_g is the number of channels) be the global feature vector produced by the final linear layer of a standard CNN classifier. Usually, $\boldsymbol{g}$ is encoded with global and dicriminative feature information of the interested object. The idea is to combine both $\boldsymbol{l}_i^s$ and $\boldsymbol{g}$ to focus on features at different spatial scales which are relevant to the high-level features represented by $\boldsymbol{g}$. For this reason, the concept of compatibility score $C\left(\mathcal{L}^s, \boldsymbol{g}\right) = \left\{c_i^s\right\}_{i=1}^n$ has been proposed, which is defined by an addictive attention model: $c_i^s = \left\langle \boldsymbol{\Psi}, \boldsymbol{l}_i^s + \boldsymbol{g} \right\rangle$, $i \in \{1, ..., n\}$, where $\langle\cdot,\cdot\rangle$ is dot product and $\boldsymbol{\Psi} \in \mathbb{R}^{C_s}$ is a learned weight vector interpreted as learning the general concept of objectness. In this case, $\boldsymbol{l}_i^s$ and $\boldsymbol{g}$ have different dimensions. Therefore, a learnable weight $W_g \in \mathbb{R}^{C_s \times C_g}$ is utilized to align the dimensionality of $\boldsymbol{g}$ to $\boldsymbol{l}_i^s$.

Then the compatibility scores are computed and passed through soft-max operation to obtain the normalized attention coefficient $\alpha_i^l = e^{c_i^l}/\sum_i e^{c_i^l}$. In the end, a weighted sum $\boldsymbol{g}^s = \sum_{i=1}^n \alpha_i^s l_i^s$ is computed, and the final prediction is given by feeding $\left\{\boldsymbol{g}^1, ..., \boldsymbol{g}^s\right\}$ to a fully connected layer. The attention coefficients α_i^l can be regarded as salient image regions which amplify the relevant information and suppress the irrelevant one so that the network is forced to learn the most discrimitive features which contribute to the final prediction to the greatest extent. Based on [12], [13] proposed a more general attention mechanism: $c_i^s = \boldsymbol{\Psi}_{\sigma_1}(W_l l_i^s + W_g \boldsymbol{g} + b_g) + b_\psi$, where bias terms $b_g \in \mathbb{R}^{C_{int}}, b_\psi \in R$, and linear transformations $\boldsymbol{\Psi} \in \mathbb{R}^{C_{int}}$, $W_l \in \mathbb{R}^{C_{int} \times C}$, $W_g \in \mathbb{R}^{C_{int} \times C_g}$ characterise a gating unit. σ_1 refers to ReLU activation function.

The introduced W_l enables the fine-scale layer focus less on its compatibility to g and concentrate more on learning discriminant features [13]. By introducing σ, the attention module is allowed to learn nonlinear relationships between these vectors. However, We find that both of the methods miss the channel-wise attention, which plays an important role in distinguishing objects.

Therefore, we come up with the idea of squeeze and excitation block (SE) [4] that makes the network to perform feature recalibration.

The SE-block acts as a computational module for any transformation $F_{tr} : X \rightarrow \tilde{X}, X \in \mathbb{R}^{H' \times W' \times C'}, \tilde{X} \in \mathbb{R}^{H \times W \times C}$. The outputs of F_{tr} are represented as $\tilde{X} = \{\tilde{x_1}, \tilde{x_2}, ..., \tilde{x_c}\}$. The whole process tries to adjust the filter responses in two steps, squeeze and excitation.

By using a global average pool, the squeeze operation extracts the contextual information from the local features to generate channel-wise statistics, $z_c \in \mathbb{R}^C$. The c-th element of z is calculated by $z_c = F_{sq}(x_c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} x_c(i, j)$.

To make use of the aggregated information from the squeeze operation, an excite operation is followed, which aims to capture channel-wise dependencies. To achieve this, a simple gating mechanism with sigmoid activation is applied: $u = F_{ex}(z, W) = \sigma(W_2 \delta(W_1, z))$, where σ is sigmoid activation and δ refers to the ReLU function, $W_2 \in \mathbb{R}^{C \times \frac{C}{r}}$ and $W_1 \in \mathbb{R}^{\frac{C}{r} \times C}$ which are used to decrease the model complexity and increase generalization.

Finally, the rescaled output of the block is represented as $\tilde{x_c} = F_{scale}(x_c, u_c) = u_c x_c$, where $\tilde{x} = \{\tilde{x_1}, \tilde{x_2}, ..., \tilde{x_c}\}$ and $F_{scale}(x_c, s_c)$ refers to channel-wise multiplication between the feature map $x_c \in \mathbb{R}^{H \times W}$ and the scalar u_c.

To combine the channel-wise and spatial-wise attention, we introduce a new g^s by: $\tilde{\mathcal{L}} = F_{tr}(\mathcal{L})$, $g^s = \sum_{i=0}^{n} \alpha^s \tilde{l_i^s}$, where F_{tr} represents the output of SE-block.

As shown above, since the global feature vector g is a 1D vector, we speculate that the flattening operation may have the disadvantage of losing important context information. Hence, we take the feature map just before the global pooling layer as instance of g. And we up-sampling the coarse scale feature map to better generate the attention maps.

For normalizing the compatibility coefficients, we believe that soft-max operation is not always the optimal choice as it typically generates sparse output, making the model over-sensitive to local changes despite we want the network to attend to the region with different parts. Thus, we adopt sigmoid function to normalize the compatibility coefficients.

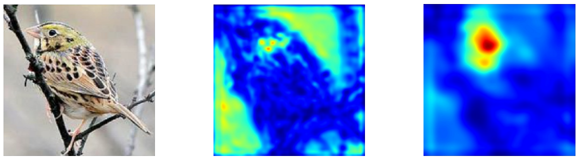

Image Attention 1 Attention 2

Figure 2: Visualization of the attention regions detected by the attention blocks. The first column (Attention 1) shows the input image, the next two columns (Attention 2) show the heat maps from the first and second attention blocks of our network respectively.

2.3 Loss Function

The binomial deviance loss [19] is used to train our aforementioned attention block.

3 EXPERIMENTS

3.1 Datasets

We evaluate our approach on four standard datasets: In-shop Clothes Retrieval [8], CUB-200-2011 (CUB) [20], Stanford Online Products (SOP) [17] and Cars-196 [7].

3.2 Implementation Details

We adopted the deep network (VGG16) of Simonyan et al. [16] pre-trained on the ImageNet ILSVRC challenge as a starting point for all of our experiments. At the step of data pre-processing, we first resize the images to 256 × 256 pixels, then we perform the data augmentation by randomly cropping the image to a size of 224 × 224 pixels and applying horizontal flipping. Each input image is normalized through mean RGB-channel subtraction. Futhermore, optimization is performed using Adam optimizer with weight decay $5x10^{-4}$ and a mini-batch size of 80. The initial learning rate is set to 10^{-5} and decreased by a factor of 10 after 40000 Iterations. The output embedding dimension is set to 512.

3.2.1 *Evaluation Measure.*

We conduct the experiments on all commonly adopted image retrieval task datasets and utilize Recall@K metric for evaluation.

3.2.2 *Ablation Setting.*

We perform experiment on the system settings include as below. Our baseline is deep network (VGG16) with the binomial deviance loss. Then the proposed ABIR without SE-block is tested. Furthermore, the whole ABIR with SE-block is used for image retrieval.

3.3 Attention Map Visualization

The attention obtained from query images are visualised in Figure 2. The first attention block appears to focus on the surroundings (see Attention 1), which combines the information from Conv3_3 and Conv5_3 layers. While the second attention block appears to focus on the object parts (see Attention 2), which combines the information from Conv4_3 and Conv5_3 layers.

3.4 State-of-the-Art Comparison

In comparison with [3, 6, 10, 22, 25] on the the CUB and Cars-196 dataset respectively, outcomes are illustrated in Table 1.

On the In-shop Clothes Retrieval dataset, we compare our model with [3, 6, 8, 10, 25]. As shown in Table 2, ABIR achieves state-of-the-art at 89.0 R@1.

And our method is compared with [3, 6, 10, 17, 25] on the Stanford Online Products (SOP) dataset. ABIR achieves nearly state-of-art results as shown in Table 3, except compared to [6] at R@1 and R@10.

3.5 Result Analysis

Quantitative experimental results on the four datasets are shown in Table 1, Table 2 and Table 3.

We first analyze the results on the CUB dataset in Table 1. It can be observed that with ABIR, our method achieves the best overall performance against state-of-the-art. It has improved significantly by 12.6 % at R@1, which demonstrates the effectiveness of our approach. Our method exhibits similar performances on the Cars-196 dataset in the meanwhile. As shown in Table 1, ABIR improves the retrieval accuracy by 2.5% at R@1.

Additionally, the In-shop Clothes Retrieval and the Stanford Online Products (SOP) datasets are more challenging since there are only few (≈ 5) images per class. We observe that it has promotion over all experiments on In-shop Clothes Retrieval dataset from Table 2, however, such extraordinary results aren't reproduced in

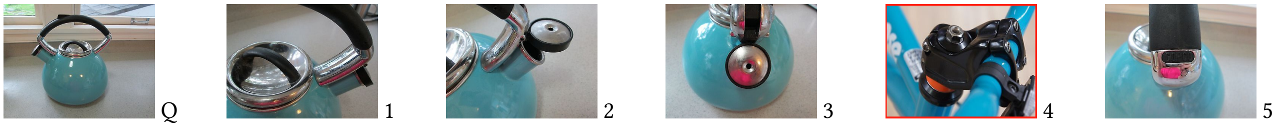

Figure 3: An example of retrieval results on the Stanford Online Products (SOP) dataset [17]. The top-5 ranking results for query (Q) are exhibited and the incorrect result are highlighted with red bounding box.

Table 1: Comparisons on the CUB-200-2011 [20] and Cars-196 [7] dataset

	CUB-200-2011						Cars-196					
R@	1	10	20	30	40	50	1	2	4	8	16	32
margin [22]	63.9	75.3	84.4	90.6	94.8	-	86.9	92.7	95.6	**97.6**	**98.7**	-
HDC [25]	60.7	72.4	81.9	89.2	93.7	96.8	83.8	89.8	93.6	96.2	97.8	98.9
HTL [3]	57.1	68.8	78.7	86.5	92.5	95.5	81.4	88.0	92.7	95.7	97.4	99.0
A-BIER [10]	65.5	75.8	83.9	90.2	94.2	**97.1**	82.0	89.0	93.2	96.1	-	-
ABE-8 [6]	60.6	71.5	79.8	87.4	-	-	85.2	90.5	94.0	96.1	-	-
Our Baseline	73.1	81.9	87.6	91.4	93.8	96.2	82.6	88.1	92.4	95.3	97.4	98.4
ABIR w/o SE-block	77.5	84.1	88.7	91.7	94.2	96.3	89.1	93.1	95.4	97.2	98.3	99.1
ABIR with SE-block	**78.1**	**84.6**	**88.7**	**91.8**	**94.4**	96.6	**89.4**	**93.3**	**95.6**	97.1	98.2	**99.0**

Table 2: Comparisons on the In-Shop Clothes Retrieval [8] dataset

R@	1	10	20	30	40	50
FashionNet+Joints [8]	41.0	64.0	68.0	71.0	73.0	73.5
FashionNet+Poselets [8]	42.0	65.0	70.0	72.0	72.0	75.0
FashionNet [8]	53.0	73.0	76.0	77.0	79.0	80.0
HDC [25]	62.1	84.9	89.0	91.2	92.3	93.1
HTL [3]	80.9	94.3	95.8	97.2	97.4	97.8
A-BIER [10]	83.1	95.1	96.9	97.5	97.8	98.0
ABE-8 [6]	87.3	96.7	97.9	98.2	98.5	98.7
Our Baseline	85.4	96.1	97.3	97.8	98.1	98.3
ABIR w/o SE-block	88.1	96.9	97.6	98.1	98.3	98.5
ABIR with SE-block	**89.0**	**97.1**	**98.0**	**98.4**	**98.6**	**98.8**

Table 3: Comparisons on the Stanford Online Products (SOP) [17] dataset

R@	1	10	100	1000
Contrastive [17]	42.0	58.2	73.8	89.1
Triplet [17]	42.1	63.5	82.5	94.8
LiftedStruct [17]	62.1	79.8	91.3	97.4
HDC [25]	69.5	84.4	92.8	97.7
HTL [3]	74.8	88.3	94.8	98.4
A-BIER [10]	74.2	86.9	94.0	97.8
ABE-8 [6]	**76.3**	**88.4**	94.8	98.2
Our Baseline	71.2	85.6	93.5	97.7
ABIR w/o SE-block	74.3	87.4	94.6	98.3
ABIR with SE-block	74.8	87.7	**95.0**	**98.5**

the other dataset, even though we also get promising results on the Stanford Online Products (SOP) dataset according to Table 3.

As shown in Figure 3, some of the wrong search results (which are highlighted with red bounding box) about Stanford Online Products (SOP) dataset are caused by the image only showing parts of the target object. This kind of picture that only shows partial information appears much more than the other three datasets, thus the difficulty of retrieval on SOP dataset is increased, which results in the improvements over three datasets at R@1 except for SOP dataset.

Furthermore, we consistently improve our strong baseline method by a large margin on all datasets, which shows the robustness of our Attention Blocks.

4 CONCLUSIONS

In this paper, we design a novel end-to-end Attention-block network for the task of Image Retrieval (ABIR). We note that our network dealt with image retrieval using multi-dimension feature fusion and a common but helpful loss function. Additionally, our approach with less number of parameters is a weakly supervised learning method, which does not require bounding box or part annotation and thus is has good generalization ability. Intuitively, the benefit of attention blocks is shown in this paper via heat maps. Extensive experiments on four datasets demonstrate that ABIR promote state-of-the-art image retrieval accuracy.

Future work interesting to us includes exploring more fast approaches to help increase the speed of image retrieval.

5 ACKNOWLEDGEMENTS

This work was supported in part by National Natural Science Foundation of China (No. U1509206).

REFERENCES

[1] Artem Babenko, Anton Slesarev, Alexander Chigorin, and Victor S. Lempitsky. 2014. Neural Codes for Image Retrieval. In *ECCV 2014*, Vol. 8689. Springer, 584–599.

[2] Charles Corbière, Hedi Ben-younes, Alexandre Ramé, and Charles Ollion. 2017. Leveraging Weakly Annotated Data for Fashion Image Retrieval and Label Prediction. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops,2017*. 2268–2274.

[3] Weifeng Ge, Weilin Huang, Dengke Dong, and Matthew R. Scott. 2018. Deep Metric Learning with Hierarchical Triplet Loss. In *ECCV 2018*, Vol. 11210. Springer, 272–288.

[4] Jie Hu, Li Shen, and Gang Sun. 2018. Squeeze-and-Excitation Networks. In *IEEE, CVPR 2018*. 7132–7141.

[5] Junshi Huang, Rogerio S. Feris, Chen Qiang, and Shuicheng Yan. 2015. Cross-domain Image Retrieval with a Dual Attribute-aware Ranking Network. In *IEEE, CVPR 2015*. 1062–1074.

[6] Wonsik Kim, Bhavya Goyal, Kunal Chawla, Jungmin Lee, and Keunjoo Kwon. 2018. Attention-based Ensemble for Deep Metric Learning. In *ECCV 2018*, Vol. 11205. Springer, 760–777.

[7] Jonathan Krause, Michael Stark, Jia Deng, and Fei Fei Li. 2014. 3D Object Representations for Fine-Grained Categorization. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2014*. 554–561.

[8] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. 2016. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *IEEE, CVPR 2016*. 1096–1104.

[9] Joe Yue-Hei Ng, Fan Yang, and Larry S. Davis. 2015. Exploiting local features from deep networks for image retrieval. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2015*. 53–61.

[10] Michael Opitz, Georg Waltner, Horst Possegger, and Horst Bischof. 2018. Deep Metric Learning with BIER: Boosting Independent Embeddings Robustly. *CoRR* abs/1801.04815 (2018). http://arxiv.org/abs/1801.04815

[11] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. 2014. CNN Features Off-the-Shelf: An Astounding Baseline for Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2014*. 512–519.

[12] Namhoon Lee Saumya Jetley, Nicholas A. Lord and Philip H. S. Torr. 2018. Learn To Pay Attention. In *International Conference of Learning Representation.*

[13] Jo Schlemper, Ozan Oktay, Michiel Schaap, Mattias P. Heinrich, Bernhard Kainz, Ben Glocker, and Daniel Rueckert. 2018. Attention Gated Networks: Learning to Leverage Salient Regions in Medical Images. *CoRR* abs/1808.08114 (2018). http://arxiv.org/abs/1808.08114

[14] Peichung Shih and Chengjun Liu. 2005. Comparative Assessment of Content-Based Face Image Retrieval in Different Color Spaces. In *2005 Audio- and Video-Based Biometric Person Authentication*, Vol. 3546. Springer, 1039–1048.

[15] Karen Simonyan and Andrew Zisserman. 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. *CoRR* abs/1409.1556 (2014). http://arxiv.org/abs/1409.1556

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

[17] Hyun Oh Song, Xiang Yu, Stefanie Jegelka, and Silvio Savarese. 2016. Deep Metric Learning via Lifted Structured Feature Embedding. In *IEEE, CVPR 2016*. 4004–4012.

[18] Giorgos Tolias, Ronan Sicre, and Hervé Jégou. 2015. Particular object retrieval with integral max-pooling of CNN activations. *CoRR* abs/1511.05879 (2015). http://arxiv.org/abs/1511.05879

[19] Evgeniya Ustinova and Victor S. Lempitsky. 2016. Learning Deep Embeddings with Histogram Loss. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016*. 4170–4178.

[20] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. 2011. *The Caltech-UCSD Birds-200-2011 Dataset*. Technical Report.

[21] Xiu-Shen Wei, Jian-Hao Luo, Jianxin Wu, and Zhi-Hua Zhou. 2017. Selective convolutional descriptor aggregation for fine-grained image retrieval. *IEEE Transactions on Image Processing* 26, 6 (2017), 2868–2881.

[22] Chao Yuan Wu, R. Manmatha, Alexander J. Smola, and Philipp KrÃďhenbÃijhl. 2017. Sampling Matters in Deep Embedding Learning. In *IEEE, CVPR 2017*. 2859–2867.

[23] Lingxi Xie, Jingdong Wang, Bo Zhang, and Qi Tian. 2015. Fine-grained image search. *IEEE Transactions on Multimedia* 17, 5 (2015), 636–647.

[24] Kota Yamaguchi, M Hadi Kiapour, and Tamara L Berg. 2013. Paper doll parsing: Retrieving similar styles to parse clothing items. In *IEEE, CVPR 2013*. 3519–3526.

[25] Yuhui Yuan, Kuiyuan Yang, and Zhang Chao. 2017. Hard-Aware Deeply Cascaded Embedding. In *IEEE, CVPR 2017*. 814–823.

[26] Wengang Zhou, Houqiang Li, and Tian Qi. 2017. Recent Advance in Content-based Image Retrieval: A Literature Survey. abs/1706.06064 (2017).