CROSS-DOMAIN BEAUTY AND PERSONAL CARE PRODUCT RETRIEVAL

[13]*Tsun-Hsien Tang (湯忠憲),* [2]*Yu-Siang Huang (黃郁翔),* [3]*Chiou-Shann Fuh (傅楸善)*

[1]Data Science Degree Program
[2]Graduate Institute of Networking and Multimedia
[3]Department of Computer Science and Information Engineering
National Taiwan University, Taipei, Taiwan
E-mail: {r06946003, r06944057}@ntu.edu.tw, fuh@csie.ntu.edu.tw

ABSTRACT

Cross-domain image retrieval is a challenging problem due to the data variations between the real-world images and advertisement images. In this work, we consider four state-of-the-art deep learning based model to extract the high-level features combining with four feature pooling strategies. Different from previous works, we further investigate the possibility of integrating the classical feature descriptors. A dataset containing half a million images of beauty and care products (Perfect-500k) is utilized for our experiments. The experimental results prove that our proposed hybrid framework can improve the mAP@7 between 3% and 10% in contrast with retrieval methods only utilizing deep features.

***Index Terms*—** cross-domain image retrieval, SIFT, deep learning, hybrid framework

1. INTRODUCTION

With the rapid growth of online shopping, the beauty and personal care merchandise are now easily available on e-commerce websites, e.g., Amazon[1] and ebay[2]. To make users access products of interest conveniently, text-based searching engines have been well-developed. However, it is hard to find a specific product without its name. Therefore, an image content-based retrieval systems, which searches items based on given product photos, is needed.

In practice, the e-commerce sites tend to demonstrate the property and appearance of products via appealing photos as shown in Figure1(a). We can treat them as reference images with query images shot by user. However, as pictures on sites have been well manipulated to enhance the quality of visual presentation on products, there might exist subtle difference from that of in real looks. As a result, the phenomena leads to a cross-domain image retrieval problem.

In order to promote impactful research and problem solving in beauty space, the leading beauty app developer, Perfect Corp. in collaborating with CyberLink Corp. and Academia Sinica to provide a large-scale image dataset of over half million images of beauty and personal care products, namely the Perfect-500K dataset [1]. Perfect-500K is vast in scale, rich and diverse in content in order to collect as many as possible beauty and personal care items from major e-commerce sites. Given a real-world image containing one beauty or personal care item, in this paper, we propose methods to match the real-world query images of item to the same items in the Perfect-500K data set. Our proposed hybrid framework leverages features generated by deep convolutional neural networks and re-ranks a short-list of retrieved image by SIFT feature matching.

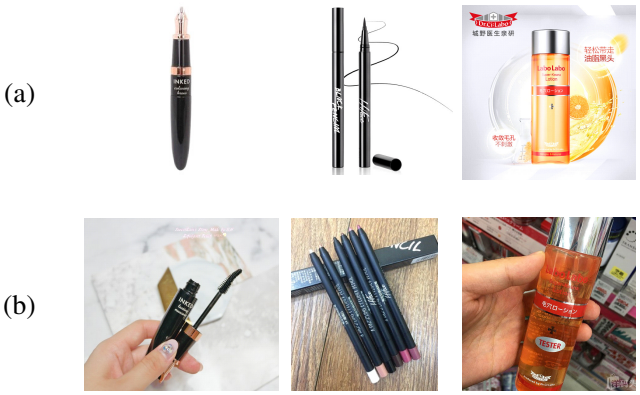

Fig. 1. Example of the cross-domain beauty and personal care product images.The buyer domain images (a) may have different viewpoints, brightness, state and complex backgrounds while the seller domain images (b) are captured in fixed and ideal conditions in clean background.

2. RELATED WORK

With the prevalence of CNN, image retrieval is also embraced with deep learning. Oquab et al. [2] proposed to use the activations of the fully connected layers as the global descriptors.

[1]https://www.amazon.com/
[2]https://www.ebay.com/

Babenko et al. [3] found that using sum-pooling to combine deep features on the last convolutional layer can obtain effective performance, and proposed the sum-pooling convolutional (SPoC) method. Tolias et al. [4] build compact feature vectors that encode several image regions without the need to feed multiple inputs to the network,in spirit of recent Fast-RCNN [5] and Faster-RCNN [6] methods but here targeting particular object retrieval.

In addition to deep learning based method, classical image descriptors serve as promising techniques for object localization as well. Arandjelovic & Zisserman [7] propose a localization strategy based on VLAD, where similarity is computed for multiple image regions, giving a more precise localization via regression. Zheng et al. [8] present milestones in modern instance retrieval, review a broad selection of previous works in different categories, and provide insights on the connection between SIFT and CNN-based methods.

3. METHODOLOGY

3.1. Classic Feature Extraction

3.1.1. Scale-Invariant Feature Transform (SIFT)

SIFT [9] is a feature detection algorithm in computer vision to detect and describe local features in images by extracting key-points and computing the corresponding descriptors. There are mainly four steps involved in the algorithm. We would simply explain them one by one.

The first step is scale-space extrema detection. Laplacian of Gaussian (LoG) is found for the image with various σ values, which acts as a blob detector and detects blobs in various sizes due to change in σ. For example, Gaussian kernel with low σ gives high value for small corner while Gaussian kernel with high σ fits well for larger corner. Therefore, we can find the local maximum across the scale and space which gives us a list of (x, y, σ) values, i.e., there is a potential key-point at (x, y) at σ scale.

Due to the procedure consumption of LoG, an alternative is to use Difference of Gaussian (DoG) for the approximation of LoG. DoG is obtained as the difference of Gaussian blurring of an image with two different σ, called σ and $k\sigma$, respectively. Once this DoG are found, images are searched for local extreme values over scale and space. A potential key-point is defined as a detected local extreme value and is best represented in that scale.

Once potential key-points locations are derived, they have to be refined to get more accurate results. In detail, Taylor series expansion of scale space is used to acquire more accurate location of extreme values. Once the intensity of a specific extreme value is less than a threshold value, it would be rejected. Moreover, a 2-by-2 Hessian matrix is applied to remove the key-points of edges and eliminates any low-contrast key-points and edge key-points. The remaining points are relatively strong interest points.

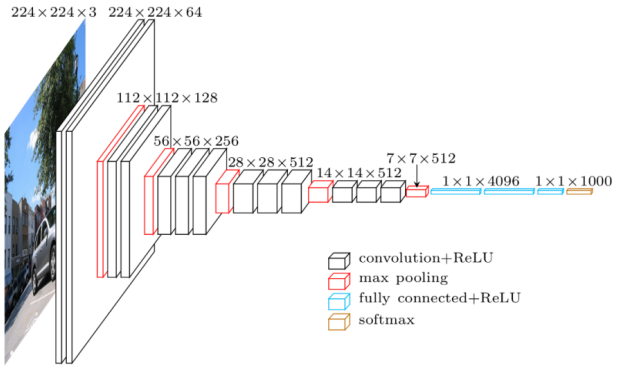

Fig. 2. Architecture of VGG16 [11].

Then, an orientation is assigned to each key-point to achieve invariance to image rotation. A neighbourhood is taken around the key-point location depending on the scale, and the gradient magnitude and direction is calculated in that region. An orientation histogram with 36 bins covering 360 degrees is created and the highest peak in the histogram is taken then any peak above 80% of it is also considered to calculate the orientation. It creates key-points with same location and scale, but different directions. It contributes to stability of matching.

Now key-point descriptor is created. A 16x16 neighbourhood around the key-point is taken. It is divided into 16 sub-blocks of 4-by-4 size. For each sub-block, 8 bin orientation histogram is created. So a total of 128 bin values are available. It is represented as a vector to form key-point descriptor.

After extracting the SIFT descriptors of all images, we apply the fast library for approximate nearest neighbors (FLANN) [10] to match the feature descriptors between two images. Key-points between two images are matched by identifying their nearest neighbours. But in some cases, the second closest-match may be very near to the first. It may happen due to noise or some other reasons. In that case, ratio of closest-distance to second-closest distance is taken. If it is greater than 0.8, they are rejected. It eliminates around 90% of false matches while discards only 5% correct matches.

For the similarity score, we first consider the number of matches between two images. Furthermore, we consider the distances between matched descriptors to avoid tie problems.

3.2. Deep Feature Extraction

In recent years, deep learning has achieved remarkable success in various artificial intelligence research areas including computer vision. Convolutional neural networks have enjoyed a great success in large-scale image and video recognition and learning efficient dense image representations. Therefore, in the case of content based image retrieval, we first extract visual representations from backbone networks and compare the similarity between the query representation

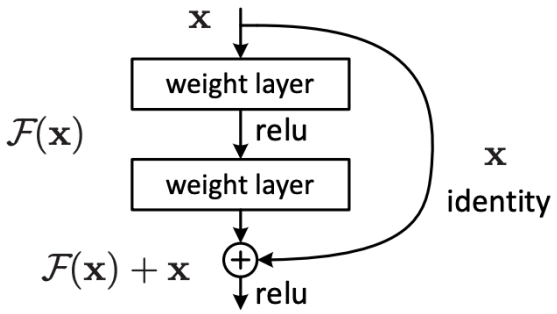

Fig. 3. Residual block in ResNet [12]

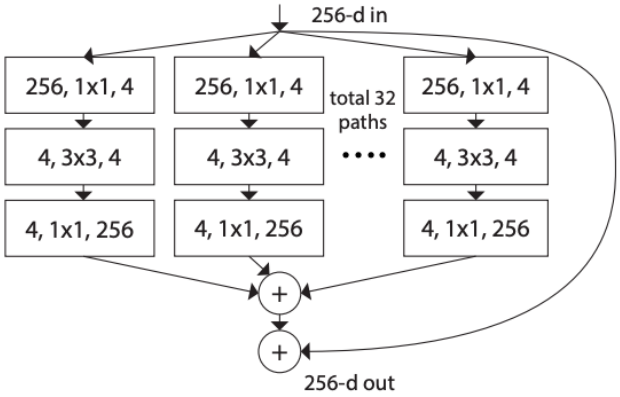

Fig. 4. A block of ResNeXt with cardinality = 32, with roughly the same complexity. A layer is shown as (# in channels, filter size, # out channels).

and representations from database. Here we leverage several powerful CNN-based image classifiers as our backbone networks for feature extraction.

3.2.1. VGG16

VGG [13] explores the use of increasing networks depth with very small (3×3) convolution filters and demonstrates that a significant improvement on the prior-art configurations can be achieved by pushing the depth to 16–19 weight layers. Here we adopt VGG with 16 layers (VGG16) and pre-training weights optimized using ImageNet dataset [14]. The architecture overview of VGG16 is shown in Figure 2. To extract image representation, we transform the features derive last convolutional layer with 512 channels by several different pooling strategies discussed in following sections.

3.2.2. ResNet

He et al. [12] present a residual learning framework to ease the training of networks that are substantially deeper than those used previously. ResNet explicitly reformulate the layers as learning residual functions with reference to the layer

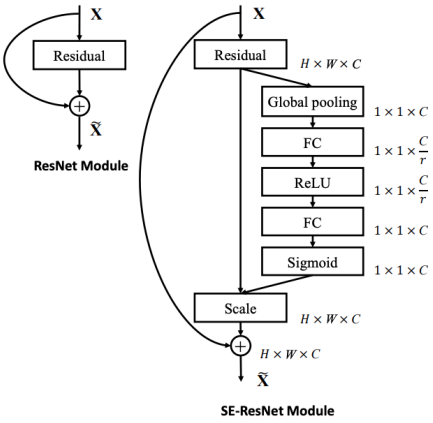

Fig. 5. The schema of the original Residual module (left) and the SE-ResNet module (right).

inputs, as shown in Figure 3, instead of learning unreferenced functions. The last convolution layers result in 2048 channels.

3.2.3. ResNeXt

Xie et al. [15] present a simple, highly modularized network architecture for image classification. The network is constructed by repeating a building block that aggregates a set of transformations with the same topology as shown in Figure 4. The design results in a homogeneous, multi-branch architecture that has only a few hyper-parameters to set. Moreover, this strategy exposes a new dimension, named "cardinality" (the size of the set of transformations), as an essential factor in addition to the dimensions of depth and width. Similar to ResNet, the final convolution layers consist of 2048 channels.

3.2.4. SENet

Hu et al. [16] propose a novel architectural unit, which called "Squeeze-and-Excitation" (SE) block, that adaptively recalibrates channel-wise feature responses by explicitly modelling inter-dependencies between channels. The work basically modifies the residual block into SE block and achieve the first place in Large Scale Visual Recognition Challenge 2017 (ILSVRC2017) [17]. The design block unit is shown in Figure 5.

3.3. Deep Feature Pooling

Instead of using feature vectors generated by the last fully connected layers, recent works derive visual representations from the activations of the convolutional layers. According to [18] such representation offers better generalization properties for test data that are far from the source (training) data.

For pooling layer, our goal is to obtain a image representation vector f given the activations (3D tensor) of a convolution layer. For further illustration, we formulate the 3D tensor of $w \times h \times d$ dimensions into a set of 2D features: $S = \{S_n\}(n = 1, ..., d)$, where S_n is the n-th channel feature of size $w \times h$.

In the following paragraphs, we introduce pooling strategies we used for image feature generations.

3.3.1. MAC

Maximum activations of convolutions (MAC) [2] encodes the maximum **local** response of each of the convolutional filters and is therefore translation invariant. The operation could be represent as

$$f_{MAC} = [f_1, f_2, ..., f_n, ..., f_d], \ f_n = \max_{x \in S_n} x. \quad (1)$$

3.3.2. SPoC

In [3], Babenko et al. investigate possible ways to aggregate local deep features to produce compact global descriptors for image retrieval. The experimental results reveal that the simple aggregation method based on sum pooling provides arguably the best performance for deep convolutional features. Thus we take sum-pooled convolutional features into consideration. In practice, we re-formulate the features as average-pooled features.

$$f_{SPoC} = [f_1, f_2, ..., f_n, ..., f_d], \ f_n = \frac{1}{|S_n|} \sum_{x \in S_n} x. \quad (2)$$

3.3.3. R-MAC

Regional maximum activation of convolutions [4] is a quite popular descriptor, which uses a multi-scale rigid grid with overlapping and generates a single feature vector per region. Here, we denote these region-level features as $R = R_k(k = 1, \ldots, m)$ for a number of m regions. The regions R are defined on the space of all valid positions for the considered feature map (and not on the input image plane). These region-level features are normalized and sum-aggregated independently to obtain the compact vectors as follows:

$$f_{R-MAC} = [f_1, f_2, ..., f_n, ..., f_d], \ f_n = \sum_{k=1}^{m} \frac{R_k}{norm(R)}. \quad (3)$$

3.3.4. RA-MAC

We can find out that R-MAC treats each region equally by summing operation and is not robust enough to deal with image from different domain (with different kind of background information). To assign larger weights to the regions containing objects instead of backgrounds, recently research [19] equips the pooling strategy with attention mechanism on top of the regions. The formulation of Regional Maximum Activations of Convolutions (RA-MAC) is shown in below.

$$M_{i,j} = \begin{cases} 1 & \text{if } A_{i,j} > \frac{\sum_i^w \sum_j^h A_{i,j}}{w \times h}, \\ 0 & otherwise \end{cases} \quad (4)$$

$$f_{RA-MAC} = [f_1, ..., f_n, ..., f_d], \ f_n = \sum_{k=1}^{m} \frac{\frac{\sum_{(i,j) \in r} M_{i,j}}{\text{size}(r)} R_k}{norm(R)}, \quad (5)$$

where $A_{i,j} = \sum_{n=1}^{d} S_n$, (i, j) is a particular position in the position of $w \times h$ and size(r) denotes the size of the region r.

3.4. Retrieval, Localization and re-ranking

3.4.1. Initial retrieval

The deep features, such as MAC, SPoC, R-MAC, RA-MAC feature vectors, are computed for all databases images. Similarly, at query time we process the query image and extract the corresponding feature vector. During the filtering stage we directly evaluate the distance between the query and all the database vectors. Therefore, we obtain the initial ranking based on the similarity of deep feature vectors. For computing the similarity (distance $\mathcal{D}$) between two feature vectors, u and v, we consider three different metrics.

- **L1 distance** [20]: the sum of the horizontal and vertical distances between points on a grid and also known as Manhattan distance.

$$\mathcal{D}_{L1} = \|u - v\| \quad (6)$$

- **L2 (Euclidean) distance** [21]: the "ordinary" straight-line distance between two points in Euclidean space.

$$\mathcal{D}_{L2} = \|u - v\|_2 \quad (7)$$

- **Cosine distance** [22]:a measure of similarity between two non-zero vectors of an inner product space that measures the cosine of the angle between them.

$$\mathcal{D}_{cos} = 1 - \frac{u \cdot v}{\|u\|_2 \|v\|_2} \quad (8)$$

3.4.2. Re-ranking

So far, a rough localization of the query object is derived from deep features. We now consider a second re-ranking stage, as typically performed in spatial verification [23] with local features. A short-list of N top-ranked images is considered and similarity measurement is applied on pairs of query and database images.

Here, we refine and re-rank our retrieval results by capturing fine-grained localization (key-points) with SIFT features given the top-100 retrieved images from database. The matching method is described in 3.1.1. We expect that the deep feature methods can help to extract relative coarse information, and SIFT features would play an important in re-ranking the results via detailed local information.

4. EXPERIMENTS

4.1. Data

Perfect-500K contains 538,517 product images collected from several e-commerce websites. To verify the retrieval performance, another 100 product images from real-world are provided as the test set. Given a real-world product image, we are asked to retrieve the same product from the Perfect-500K.

4.2. Performance Metric

For evaluating the performance, we use mean average precision with top-7 retrieved results (mAP@7) which computes the average precision (AP) for each individual query, and then compute the mean among all the queries.

$$\text{mAP@}Q = \frac{\sum_{q=1}^{Q} \text{AP}(q)}{Q} \tag{9}$$

$$\text{AP}(q) = \frac{1}{\text{GTP}} \sum_{i=1}^{q} \frac{\text{TP seen}}{i} \tag{10}$$

where Q is the number of considered retrieved results (i.e., 7 in our experiments). GTP refers to the total number of true positive samples for the query, and TP seen refers to the number of true positive samples seen till q. The higher mAP value indicates the better retrieval performance.

4.3. Evaluation Result

The evaluation results of VGGNet are in Table 1, ResNet in Table 2, ResNeXt in Table 3, and SENet in Table 4. Without the SIFT re-ranking, the best mAP@7 is 0.3329 from the SENet features with RA-MAC and L2/cosine distance. Moreover, with the SIFT re-ranking, the best mAP@7 is 0.3787 from the SENet features with MAC and cosine distance. We can see the effectiveness of the SIFT re-ranking which improves all the combination methods between 3% and 10%.

Both the best performance with or without the SIFT re-ranking are from the SENet features, which implies the features from SENet are more powerful than other models. For the different pooling methods, there is no obvious difference between MAC, R-MAC, and RA-MAC. However, the results of SPoC are generally worse than the other three pooling methods. For the different distance metrics, there is also no obvious difference between the three metrics.

Table 1. mAP@7 of VGGNet with or without the SIFT re-ranking.

	SIFT	MAC	R-MAC	RA-MAC	SPoC
L1	w/o	0.2260	0.1917	0.2040	0.1745
	w/	0.2517	0.2317	0.2600	0.2387
L2	w/o	0.2229	0.1882	0.2187	0.1783
	w/	0.2737	0.2478	0.2875	0.2208
Cosine	w/o	0.2229	0.1882	0.2187	0.1783
	w/	0.2701	0.2415	**0.2878**	0.2162

Table 2. mAP@7 of ResNet with or without the SIFT re-ranking.

	SIFT	MAC	R-MAC	RA-MAC	SPoC
L1	w/o	0.2834	0.2624	0.2938	0.2517
	w/	**0.3762**	0.3433	0.3700	0.2942
L2	w/o	0.2822	0.2774	0.2961	0.2327
	w/	0.3545	0.3392	0.3631	0.2573
Cosine	w/o	0.2822	0.2774	0.2961	0.2327
	w/	0.3603	0.3392	0.3685	0.2637

Table 3. mAP@7 of ResNeXt with or without the SIFT re-ranking.

	SIFT	MAC	R-MAC	RA-MAC	SPoC
L1	w/o	0.3059	0.2709	0.2827	0.2552
	w/	0.3675	0.3515	0.3423	0.3128
L2	w/o	0.3046	0.2648	0.2750	0.2210
	w/	0.3658	0.3550	0.3517	0.2945
Cosine	w/o	0.3046	0.2648	0.2750	0.2210
	w/	**0.3725**	0.3600	0.3517	0.2942

Table 4. mAP@7 of SENet with or without the SIFT re-ranking.

	SIFT	MAC	R-MAC	RA-MAC	SPoC
L1	w/o	0.2786	0.2853	0.2809	0.2629
	w/	0.3488	0.3301	0.3623	0.3304
L2	w/o	0.2923	0.2964	0.3329	0.2674
	w/	0.3728	0.3401	0.3653	0.3010
Cosine	w/o	0.2923	0.2964	0.3329	0.2674
	w/	**0.3787**	0.3406	0.3703	0.3001

4.4. Visualization

We demonstrate two retrieval results as shown in Tables 5 and 6 using the same backbone network (SENet) and the same distance measurement (cosine distance), respectively, with or without the SIFT re-ranking. Both of the results reveal the effectiveness of SIFT re-ranking method. Moreover, the false

	Query	Top-1	Top-2	Top-3	Top-4	Top-5	Top-6	Top-7
w/o SIFT re-ranking								
MAC								
R-MAC								
RA-MAC								
SPoC								
w/ SIFT re-ranking								
MAC								
R-MAC								
RA-MAC								
SPoC								

Table 5. An example retrieval top-7 results ("v000146.jpg" in the test set) under the SENet features and cosine distance. The true retrieved ones are highlighted by green squares.

positive samples are quite similar to the query examples, indicating the difficulty of the current task.

5. CONCLUSION

In this paper, we propose a hybrid framework by combining the deep learning features (i.e., VGGNet, ResNet, ResNeXt, SENet) with different pooling methods (i.e., MAC, R-MAC, RA-MAC, SPoC) and the classical features (i.e., SIFT). The experimental results of the beauty product retrieval task demonstrate the effectiveness of our proposed approach. However, this kind of cross-domain images retrieval task is still challenging for the real-world applications. In the future work, we plan to combine more different feature

	Query	Top-1	Top-2	Top-3	Top-4	Top-5	Top-6	Top-7
w/o SIFT re-ranking								
VGGNet								
ResNet								
ResNeXt								
SENet								
w/ SIFT re-ranking								
VGGNet								
ResNet								
ResNeXt								
SENet								

Table 6. An example retrieval top-7 results ("v000105.jpg" in the test set) under MAC pooling method and cosine distance. The true retrieved ones are highlighted by green squares.

extraction methods like using Optical Character Recognition (OCR) to recognize the product description or more powerful features (e.g., Speeded Up Robust Features (SURF)).

REFERENCES

[1] Wen-Huang Cheng, Jia Jia, Si Liu, Jianlong Fu, Johnny Tseng, and Jau Huang, "Perfect corp. challenge 2018: Half million beauty product image recognition," `https://challenge2018.perfectcorp.com/index.html`, 2018.

[2] Maxime Oquab, Leon Bottou, Ivan Laptev, and Josef Sivic, "Learning and transferring mid-level image representations using convolutional neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1717–1724.

[3] Artem Babenko and Victor Lempitsky, "Aggregating local deep features for image retrieval," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1269–1277.

[4] Giorgos Tolias, Ronan Sicre, and Hervé Jégou, "Particular object retrieval with integral max-pooling of cnn activations," *arXiv preprint arXiv:1511.05879*, 2015.

[5] Ross Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.

[6] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems (NIPS)*, 2015.

[7] Relja Arandjelovic and Andrew Zisserman, "All about vlad," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2013, pp. 1578–1585.

[8] Liang Zheng, Yi Yang, and Qi Tian, "Sift meets cnn: A decade survey of instance retrieval," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 5, pp. 1224–1244, 2017.

[9] David G Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[10] Marius Muja and David G Lowe, "Fast approximate nearest neighbors with automatic algorithm configuration.," *VISAPP (1)*, vol. 2, no. 331-340, pp. 2, 2009.

[11] "A brief report of the heuritech deep learning meetup #5," https://blog.heuritech.com/2016/02/29/a-brief-report-of-the-heuritech-deep-learning-meetup-5/.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

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

[14] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[15] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He, "Aggregated residual transformations for deep neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1492–1500.

[16] Jie Hu, Li Shen, and Gang Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[17] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.

[18] Hossein Azizpour, Ali Sharif Razavian, Josephine Sullivan, Atsuto Maki, and Stefan Carlsson, "From generic to specific deep representations for visual recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2015, pp. 36–45.

[19] Zehang Lin, Zhenguo Yang, Feitao Huang, and Junhong Chen, "Regional maximum activations of convolutions with attention for cross-domain beauty and personal care product retrieval," in *2018 ACM Multimedia Conference on Multimedia Conference*. ACM, 2018, pp. 2073–2077.

[20] Eugene F Krause, *Taxicab geometry: An adventure in non-Euclidean geometry*, Courier Corporation, 1986.

[21] Michel Marie Deza and Elena Deza, "Encyclopedia of distances," in *Encyclopedia of distances*, pp. 1–583. Springer, 2009.

[22] Pang-Ning Tan, *Introduction to data mining*, Pearson Education India, 2018.

[23] James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in *2007 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2007, pp. 1–8.