Attention-driven Unsupervised Image Retrieval for Beauty Products with Visual and Textual Clues

Jingwen Hou[*][†]
Nanyang Technological University
Singapore
jingwen003@e.ntu.edu.sg

Sijie Ji[†]
Nanyang Technological University
Singapore
sijie001@e.ntu.edu.sg

Annan Wang
Nanyang Technological University
Singapore
c190190@e.ntu.edu.sg

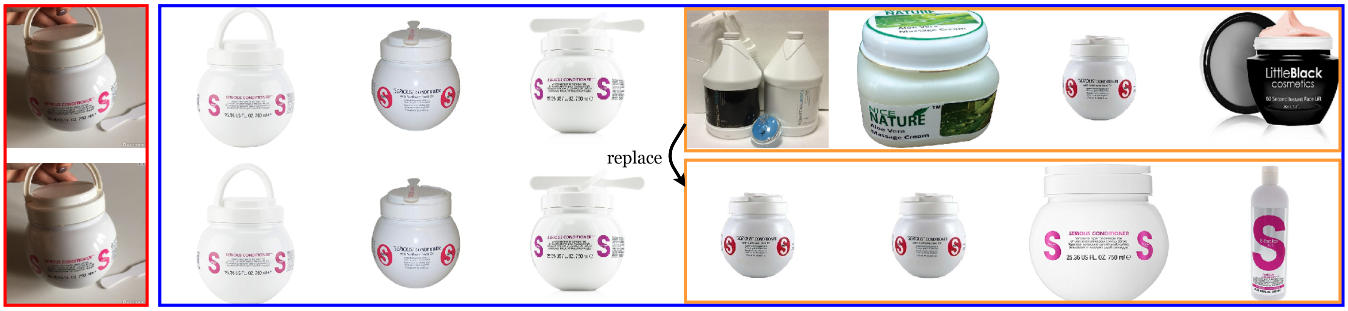
replace
Query image
Top 7 matched examples before (top) and after (bottom) the proposed refinement

Figure 1: Search results before (top) and after (bottom) the proposed refinement. The proposed refinement strategy runs a second search within the examples with product descriptions similar to the top 3 matched examples of the first search result and replaces the last 4 matched examples of the first search result with the top 4 matched examples of the second search result.

ABSTRACT

Beauty and personal care product retrieval (BPCR) aims to match a query image of an item to examples of the same item in a large database. The task is extremely challenging because a small number of ground-truth examples have to be found in a large search space. Previous works mostly search only with visual representations and have not made full use of the product descriptions. Since many noisy examples only have subtle visual differences comparing to the ground-truth examples (e.g. similar packaging but different brands) and those differences (e.g. product brands) are especially hard to be captured only by visual features, methods merely based on visual feature similarities can easily regard those noisy examples as examples of the same item in the query image. We notice that the product descriptions are good sources for capturing those subtle visual differences. Therefore, we propose a search method utilizing both images and product descriptions in this work. Before searching, we not only prepare attention-based visual features for each database image but also a textual index (TI) that matches each database example to other examples with similar product descriptions. During searching, the visual feature of the query image is firstly searched in the whole database and then searched in a subset obtained by looking up the TI. Finally, the second result is used to refine the initial result. Since the subset examples usually have similar properties (e.g. brands and type), the noisy examples in the initial result can be effectively replaced. We have experimentally proved the effectiveness of the proposed method on the validation set of the Perfect-500K dataset. Our team (NTU-Beauty) achieved the 3rd place in the leader board of the Grand Challenge of AI Meets Beauty in ACM Multimedia 2020. Our code is available at: https://github.com/jingwenh/2020-ai-meets-beauty_ntubeauty.git.

CCS CONCEPTS

• **Information systems** → **Image search**; • **Computing methodologies** → **Visual content-based indexing and retrieval**; **Image representations**.

KEYWORDS

image retrieval, attention mechanism, unsupervised learning

ACM Reference Format:
Jingwen Hou, Sijie Ji, and Annan Wang. 2020. Attention-driven Unsupervised Image Retrieval for Beauty Products with Visual and Textual Clues. In *Proceedings of the 28th ACM International Conference on Multimedia (MM '20), October 12–16, 2020, Seattle, WA, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3394171.3416271

[*] Corresponding author
[†] These authors contributed equally to this work and share first authorship.

MM '20, October 12–16, 2020, Seattle, WA, USA
© 2020 Association for Computing Machinery.
ACM ISBN 978-1-4503-7988-5/20/10…$15.00
https://doi.org/10.1145/3394171.3416271

1 INTRODUCTION

Given a real-world query image, beauty and personal care product retrieval (BPCR) aims to match the image to the examples of the same item in a large database, and each example in the database is given as an image with a piece of product description. This is a practical need desired by present online shopping platforms to enhance user experience. In this work, we attempt to propose a better solution for a BPCR challenge, *the Grand Challenge of AI Meets Beauty in ACM Multimedia 2020*. The challenge prepares a database with approximately 500k images and their product descriptions, namely Perfect-500K dataset [3]. When a real-world query image is given, the proposed method should provide the IDs of the top 7 similar database examples.

The BPCR challenge has been treated as an image retrieval task, which aims to find the most visually similar examples in the database according to the given image. A common solution is to represent each image by a feature vector and compute similarities between the features of query image and database images. The features can involve low-level components [2, 14, 19, 22, 23] such as representations for color, texture, and shape, and high-level components [1, 5, 10, 15, 16] such as CNN-based features.

Different from other image retrieval tasks, the BPCR challenge has several specific practical issues. First, the database has not provided exact labels for each product in the database. The database only provides rough descriptions of the product crawled from e-commerce websites, which contains brands, names, and specifications of the corresponding items. Since this description varies from example to example, the description cannot be directly used for training a specialized classification model. Second, the database contains images with both pure color background and noisy background (i.e. real-world images). This makes the extracted visual features can focus on objects in the background instead of the target item. Third, for beauty and personal care products, many different items have similar visual properties. For example, two lipsticks of a similar color and packaging can have different brands.

As attention mechanism and network pretraining in deep learning has greatly driven the progress in many computer vision areas [6, 7], previous methods for BPCR also tried to cope with aforementioned issues with similar approaches. Although the noisy background has been relieved by attention mechanism [13, 20] or salient object detection [11, 17] and the lack of exact labels for training have been alleviated by unsupervised learning [12, 17, 21] or CNN model pretrained on other datasets [11, 13, 18, 20], the problem brought by subtle visual differences is far from being solved since those differences are hard to be captured by visual features extracted with non-specialized CNN. Fig.1 shows an example of this problem. The search result in the first row is obtained by the top 1 solution of the 2019 Challenge [20]. Though the first 3 matched examples are very accurate, obviously irrelevant examples appear in the remaining ones. **Thus, the first intuition for improving previous solutions is to utilize the accurately matched examples to refine the remaining result.** We also notice that the product descriptions contain abundant information hard to be captured by visual representations. **This inspires us to utilize the product descriptions to capture those subtle visual differences among different items to make up for the potential inability of visual representations.**

Therefore, we propose a method that utilizes both images and product descriptions for the BPCR task. Specifically, the search is done in two stages: an initialization stage that generates the initial search result with the method [20]; and a refinement stage which refines the initial search result with the top k matched examples of the initial result and the product descriptions. To this end, before searching, attention-based visual features are extracted from all database images with a pretrained CNN model, and a textual index (TI) which indicates the most similar examples of each example in the database is generated from all product descriptions of the database. In the refinement stage after the initial result is generated, the top k examples of the initial result are found in the TI, and their similar examples are collected to form a subset. Then we search the query image in the subset. The final result is obtained by replacing the last $7 - k$ examples of the initial result with the top $7 - k$ examples of the second search result. The second row of Fig.1 shows the result after refinement. Because most examples in the subset have the same brand, type, and properties as the top 3 ($k = 3$) examples, the inaccurate examples in the initial result are effectively replaced by examples textually similar to the top 3 matched examples in the initial result.

In summary, the main contribution of this work is that we have proposed a search result refinement strategy with product descriptions that can be built upon any visual feature based searching framework. The effectiveness of our method has been verified via extensive ablation study on the Perfect-500K database, and we have achieved the 3rd place in the leader board of the Grand Challenge of AI Meets Beauty in ACM Multimedia 2020.

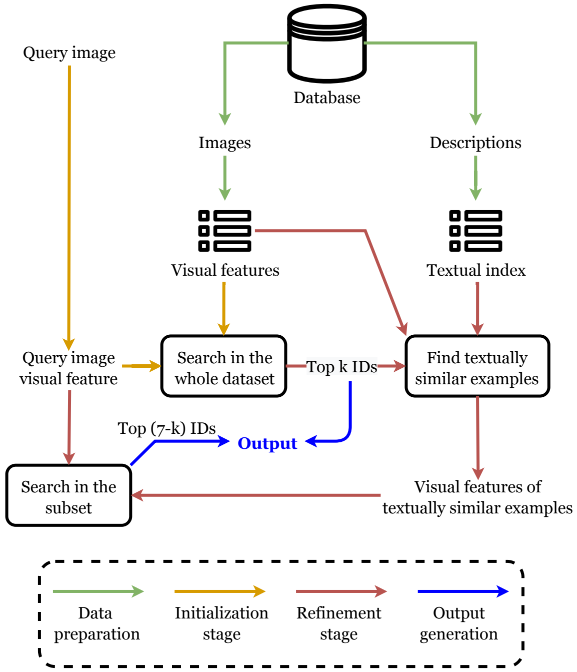

Figure 2: Overview of the proposed method. Steps of the proposed method are sequentially shown in the legend.

2 METHOD

We propose a two-stage searching method for the BPCR task. An overview of the proposed method is shown in Fig.2. Before searching, attention-based visual features are extracted from all database images (Sec.2.1), and a textual index (TI) is constructed from database product descriptions (Sec.2.2). During searching, the results are obtained from two stages (Sec.2.3), including an initialization stage and a refinement stage. Finally, the results of the first and second stages are merged for the ultimate outputs.

2.1 Attention-based visual feature extraction

We mainly follow the top 1 solution of the 2019 challenge [20] for extracting visual features. We briefly introduce the method in this section. The key idea is to zero-out the feature activation of the image background by a threshold computed from overall activation. This can be interpreted as using weakly-supervised attention to filter out noisy activation on CNN outputs. For a set of features extracted from a pretrained CNN backbone $X = \{x^k\}_{k=1}^{c}, x^k \in \mathbf{R}^{h \times w}$, we firstly take their average across channel $\tilde{X} = \sum_{k=1}^{c} x_k / c$ to obtain an average feature map $\tilde{X} \in \mathbf{R}^{h \times w}$ subject to $x_k \in X$. Then we derive the threshold t from the average feature map $\tilde{X}$:

$$t = \left(\frac{1}{wh} \sum_{i=1}^{w} \sum_{j=1}^{h} (\tilde{X}_{i,j} - \min \tilde{X})^p \right)^{\frac{1}{p}} + \min \tilde{X}, \tag{1}$$

where $\min \tilde{X}$ is the minimum value of $\tilde{X}$ and p is a parameter that controls how much the background activation is taken into consideration. As shown in [20], a larger p value corresponds to less considerations on the background. In our case, we take $p = 1$ as [20], and therefore the threshold equals the average value of $\tilde{X}$. The attention-based visual feature $\mathbf{v} \in \mathbf{R}^c$ is obtained from a set of local regions $R_s \in R$ sampled from the spatial dimension of the feature maps X:

$$\mathbf{v} = \sum_{R_s \in R} \left(\frac{\sum_{i=1}^{w_s} \sum_{j=1}^{h_s} \mathbf{1}_{\{\tilde{X}_{i,j} > t\}}}{w_s h_s} GMP(X_{R_s}) \right), \tag{2}$$

where $\mathbf{1}_{\{\tilde{X}_{i,j} > t\}}$ is an indicator function, $X_{R_s} = \{x_{R_s}^k\}_{k=1}^{c}$ with $x_{R_s}^k \in \mathbf{R}^{h_s \times w_s}$ is the feature map set spatially sampled from local region R_s of X and $GMP(\cdot)$ is the global max pooling that takes the max value of each feature map of local feature set X_{R_s}.

2.2 Textual index (TI) construction

We construct the TI from product descriptions. For preprocessing, we firstly drop all non-English descriptions for further processing. Though we may translate all non-English descriptions into English, we do not translate them since some key information especially brands can be misinterpreted. Then, the words are converted to lowercase and all punctuations and other special symbols are removed. Finally, each of the preprocessed product descriptions is tokenized into a list of terms. We have not stemmed each word because words of product brands could be destroyed by the stemming process. Then each list of words in the collection of product descriptions is converted to a TF-IDF vector. The TF-IDF based weight $w_{t,d}$ between a term t and a document (product description of one example) d is computed by:

$$w_{t,d} = TF_{t,d} \times IDF_t, \tag{3}$$

$$IDF_t = \log \frac{1 + N}{1 + DF_t} + 1, \tag{4}$$

$$\mathbf{w}_d = [w_{t_1,d}, w_{t_2,d}, ..., w_{t_M,d}], \tag{5}$$

where N is the number of documents, $\mathbf{w}_d$ is the TF-IDF vector of document d and $\{t_i\}_{i=1}^{M}$ is the set of unique terms across the collection. To construct the TI, for each document, we find its similar documents by cosine similarity $s_{textual}$ across the collection:

$$s_{textual}(\mathbf{w}_{query}, \mathbf{w}_{data}) = \frac{\mathbf{w}_{query} \cdot \mathbf{w}_{data}}{\|\mathbf{w}_{query}\|_2 \|\mathbf{w}_{data}\|_2}. \tag{6}$$

Then the TI is constructed as a hash map, whose keys are example IDs (500k in total) and values are lists of example IDs of similar examples. For convenience, we denote the hash map as a function:

$$\mathcal{D}(i) = \{j | s_{textual}(\mathbf{w}_{d_i}, \mathbf{w}_{d_j}) > \tau, i \in \mathcal{A}, j \in \mathcal{A}, i \neq j\}, \tag{7}$$

where i, j are two example IDs that belong to the set of all database example IDs $\mathcal{A}$, and τ is the threshold for selecting similar examples. We set $\tau = 0.6$ in our implementation.

Algorithm 1: Two-stage searching

```
Input   : Query visual feature $\mathbf{v}_{query}$, textual index $\mathcal{D}(\cdot)$,
          database features $\mathbf{V}_{\mathcal{A}}$, parameter $k$
Output: Matched IDs $\mathcal{I}_{output} = \{j_m\}_{m=1}^{7}$
$\mathcal{I}_{query,\mathcal{A}} \leftarrow argsortSimilaritiesDscd(\mathbf{v}_{query}, \mathbf{V}_{\mathcal{A}})$;
Initialize an empty list $\mathcal{I}_{subset}$;
for $n \leftarrow 1$ to $k$ do
    Add all elements of $\mathcal{D}(\mathcal{I}_{query,\mathcal{A}}^{n})$ to $\mathcal{I}_{subset}$;
end
$\mathbf{V}_{subset} \leftarrow findFeaturesByIDs(\mathcal{I}_{subset}, \mathbf{V}_{\mathcal{A}})$;
$\mathcal{I}_{query,subset} \leftarrow argsortSimilaritiesDscd(\mathbf{v}_{query}, \mathbf{V}_{subset})$ ;
Initialize an empty list $\mathcal{I}_{output}$ ;
for $n \leftarrow 1$ to 7 do
    if $n <= k$ then
        Add $\mathcal{I}_{query,\mathcal{A}}^{n}$ to $\mathcal{I}_{output}$
    else
        Add $\mathcal{I}_{query,subset}^{n-k}$ to $\mathcal{I}_{output}$
    end
end
```

2.3 Two-stage searching

After visual features and the TI are prepared, we find the top 7 matched examples of a given query image in two searching stages. The first initialization stage searches the query image in the whole dataset to produce an initial result with 7 matched examples, while the second refinement stage searches within a subset that is found by searching the top k examples in the TI. Both stages look for matched examples of the query image by computing similarities between the visual features of the query image and database images. Given the visual features of a query image and a database image $\mathbf{v}_{query}, \mathbf{v}_{data}$, their similarity $s_{visual}(\mathbf{v}_{query}, \mathbf{v}_{data})$ is also computed by cosine similarity similar to Eq.(6). Since multiple

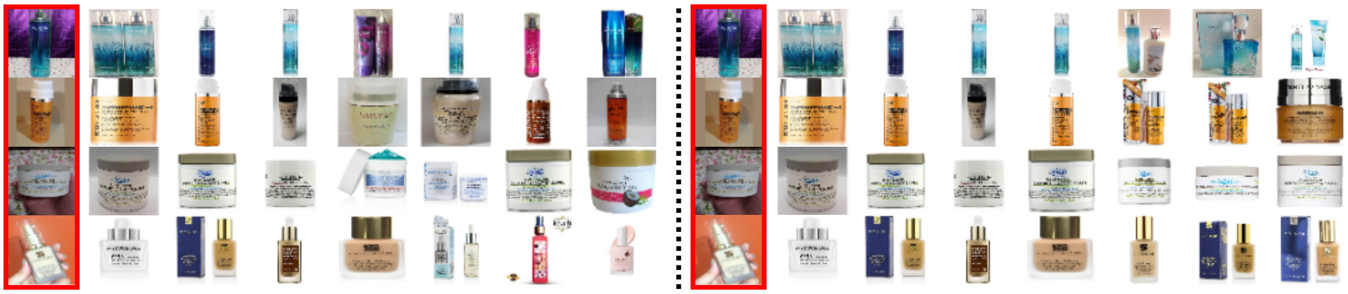

Figure 3: Initial results (left) and refined results (right) when k=3. Images in the red boxes are query images.

CNNs can be adopted, we can have a set of visual similarities $S_{visual} = \{s^n_{visual}(\mathbf{v}^n_{query}, \mathbf{v}^n_{data})\}^N_{n=1}$ resulted from N sets of visual features extracted from N different CNNs for any given query image. Different from [20] which combines the similarities by linear addition, we fuse the N visual similarities by:

$$\hat{s}_{visual} = \prod_{n=1}^{N} (1 + \alpha s^n_{visual}), \tag{8}$$

which allows similarities from each set of features to have a contribution to the overall similarities, and the strength of their contributions can be controlled by the parameter α. In our implementation, we use DenseNet201 [9] and SE-ResNet152 [8] pretrained on ImageNet [4] for extracting visual features, and values of α are empirically set to 3 and 1. The details of the searching are given as Algorithm 1. Given the query visual feature $\mathbf{v}_{query}$, TI $\mathcal{D}(\cdot)$, database visual features $\mathbf{V}_{\mathcal{A}}$ and the parameter k, the algorithm outputs top 7 example IDs $\mathcal{I}_{output}$. First, the cosine similarities between $\mathbf{v}_{query}$ and all features in $\mathbf{V}_{\mathcal{A}}$ are computed and the corresponding example IDs $\mathcal{I}_{query,\mathcal{A}}$ are arranged in descending order by the cosine similarities. Second, for top k examples in $\mathcal{I}_{query,\mathcal{A}}$, IDs of their similar examples $\mathcal{I}_{subset}$ are found in TI $\mathcal{D}(\cdot)$. Third, the cosine similarities between $\mathbf{v}_{query}$ and all features in $\mathbf{V}_{subset}$ (visual features of examples in $\mathcal{I}_{subset}$) are computed and the corresponding example IDs $\mathcal{I}_{query,subset}$ are arranged in descending order by the cosine similarities. Finally, the top k IDs in $\mathcal{I}_{query,\mathcal{A}}$ and top $7 - k$ IDs in $\mathcal{I}_{query,subset}$ are merged to form $\mathcal{I}_{output}$. $k \in [1, 7]$ controls the number of examples remained from the initial result and thus the refinement stage has no effect when $k = 7$.

3 EXPERIMENTS

3.1 Implementation details

The proposed method is evaluated on the Perfect-500K dataset. The Perfect-500K contains about 500k beauty and personal care product images collected from e-Commerce websites. Apart from the images, each image also goes with a piece of description that contains the product name, type, brand, and specification (e.g. volume and color). Since the product descriptions are also collected from e-Commerce websites, the product descriptions cannot be exactly the same for examples of the same product. A validation set with 100 real-world images of beauty and personal care products is also provided. For each example in the validation set, the IDs of ground truth matched examples in the Perfect-500K are given. The testing set is kept secret for the fairness of competition. Mean average precision of the top 7 matched examples (mAP@7) is adopted as the evaluation metric. We conduct an ablation study on the validation set before uploading our model for official testing. All images are resized to 480 × 480 and normalized before extracting visual features.

Table 1: Results of ablation study on the validation set.

Method	mAP@7	Improved %	Impaired %
Baseline	0.396944	-	-
Refined, k=6	0.397659	3	2
Refined, k=5	0.400262	6	2
Refined, k=4	0.405986	9	3
Refined, k=3	**0.407997**	**11**	**4**
Refined, k=2	0.402885	10	8
Refined, k=1	0.397293	7	6

3.2 Ablation study

The ablation study mainly investigates the effectiveness of the refinement stage with different values of parameter k. The baseline model directly takes the results from the initialization stage for evaluation. Then we assess the results from two-stage searching with different k values (1 to 6). The results are shown in Table 1. As the results shown, the refinement step can effectively improve the search results and the refinement achieves the best results when $k = 3$, which means the last 4 examples of any initial result are replaced. Note that the refinement step does not guarantee improvement on every initial result, because the TI itself does not guarantee textually matched examples are all of the same items. Thus, apart from mAP@7, we also count the proportion of results that have been improved and impaired after the refinement step. We also observe at most 11% results are improved when k=3 with 4% results are impaired. We also present some qualitative results in Fig. 3 when $k = 3$. We can see the last 4 images of the initial results are effectively replaced with more relevant ones.

4 CONCLUSION

In this paper, we have proposed a search result refinement strategy with product descriptions for BPCR that can be built upon any visual feature based searching framework. Extensive experimental analyses on the Perfect-500K validation set demonstrated the refinement strategy can effectively improve the overall performance of searching. The proposed framework achieved 3rd place of the Grand Challenge of AI Meets Beauty in ACM Multimedia 2020.

REFERENCES

[1] Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. 2016. NetVLAD: CNN architecture for weakly supervised place recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5297–5307.

[2] Anna Bosch, Andrew Zisserman, and Xavier Munoz. 2007. Image classification using random forests and ferns. In *2007 IEEE 11th international conference on computer vision*. Ieee, 1–8.

[3] Wen-Huang Cheng, Jia Jia, Si Liu, Jianlong Fu, Jiaying Liu, Shintami Chusnul Hidayati, Johnny Tseng, and Jau Huang. 2020. Perfect Corp. Challenge 2020: Half Million Beauty Product Image Recognition. https://challenge2020.perfectcorp.com/.

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 248–255.

[5] Thanh-Toan Do, Khoa Le, Tuan Hoang, Huu Le, Tam V Nguyen, and Ngai-Man Cheung. 2019. Simultaneous feature aggregating and hashing for compact binary code learning. *IEEE Transactions on Image Processing* 28, 10 (2019), 4954–4969.

[6] Jingwen Hou, Weisi Lin, and Baoquan Zhao. 2020. Content-dependency Reduction with Multi-task Learning in Blind Stitched Panoramic Image Quality Assessment. In *2020 IEEE International Conference on Image Processing (ICIP)*. IEEE.

[7] Jingwen Hou, Sheng Yang, and Weisi Lin. 2020. Object-level Attention for Aesthetic Rating Distribution Prediction. In *Proceedings of the 28th ACM International Conference on Multimedia*.

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

[9] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4700–4708.

[10] Yang Li, Yulong Xu, Jiabao Wang, Zhuang Miao, and Yafei Zhang. 2017. Ms-rmac: Multiscale regional maximum activation of convolutions for image retrieval. *IEEE Signal Processing Letters* 24, 5 (2017), 609–613.

[11] Jian Han Lim, Nurul Japar, Chun Chet Ng, and Chee Seng Chan. 2018. Unprecedented usage of pre-trained CNNs on beauty product. In *Proceedings of the 26th ACM international conference on Multimedia*. 2068–2072.

[12] Zehang Lin, Haoran Xie, Peipei Kang, Zhenguo Yang, Wenyin Liu, and Qing Li. 2019. Cross-domain Beauty Item Retrieval via Unsupervised Embedding Learning. In *Proceedings of the 27th ACM International Conference on Multimedia*. 2543–2547.

[13] Zehang Lin, Zhenguo Yang, Feitao Huang, and Junhong Chen. 2018. Regional maximum activations of convolutions with attention for cross-domain beauty and personal care product retrieval. In *Proceedings of the 26th ACM international conference on Multimedia*. 2073–2077.

[14] Krystian Mikolajczyk and Cordelia Schmid. 2004. Scale & affine invariant interest point detectors. *International journal of computer vision* 60, 1 (2004), 63–86.

[15] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. 2014. CNN features off-the-shelf: an astounding baseline for recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. 806–813.

[16] Giorgos Tolias, Ronan Sicre, and Hervé Jégou. 2015. Particular object retrieval with integral max-pooling of CNN activations. *arXiv preprint arXiv:1511.05879* (2015).

[17] Jiawei Wang, Shuai Zhu, Jiao Xu, and Da Cao. 2019. The Retrieval of the Beautiful: Self-Supervised Salient Object Detection for Beauty Product Retrieval. In *Proceedings of the 27th ACM International Conference on Multimedia*. 2548–2552.

[18] Qi Wang, Jingxiang Lai, Kai Xu, Wenyin Liu, and Liang Lei. 2018. Beauty product image retrieval based on multi-feature fusion and feature aggregation. In *Proceedings of the 26th ACM international conference on Multimedia*. 2063–2067.

[19] Xi Yang, Xinbo Gao, and Qi Tian. 2015. Polar embedding for aurora image retrieval. *IEEE Transactions on Image Processing* 24, 11 (2015), 3332–3344.

[20] Jun Yu, Guochen Xie, Mengyan Li, Haonian Xie, and Lingyun Yu. 2019. Beauty Product Retrieval Based on Regional Maximum Activation of Convolutions with Generalized Attention. In *Proceedings of the 27th ACM International Conference on Multimedia*. 2553–2557.

[21] Yi Zhang, Linzi Qu, Lihuo He, Wen Lu, and Xinbo Gao. 2019. Beauty Aware Network: An Unsupervised Method for Makeup Product Retrieval. In *Proceedings of the 27th ACM International Conference on Multimedia*. 2558–2562.

[22] Liang Zheng, Yi Yang, and Qi Tian. 2017. SIFT meets CNN: A decade survey of instance retrieval. *IEEE transactions on pattern analysis and machine intelligence* 40, 5 (2017), 1224–1244.

[23] Wengang Zhou, Houqiang Li, and Qi Tian. 2017. Recent advance in content-based image retrieval: A literature survey. *arXiv preprint arXiv:1706.06064* (2017).