

# Attention-based Pyramid Aggregation Network for Visual Place Recognition

Yingying Zhu

College of Compute Science and Software Engineering,  
Shenzhen University, China  
zhuyy@szu.edu.cn

Lingxi Xie

Department of Computer Science, Johns Hopkins  
University, USA  
198808xc@gmail.com

Jiong Wang

College of Compute Science and Software Engineering,  
Shenzhen University, China  
liubinggunzu@gmail.com

Liang Zheng

Research School of Computer Science, Australian National  
University, AUS  
liangzheng06@gmail.com

## ABSTRACT

Visual place recognition is challenging in the urban environment and is usually viewed as a large scale image retrieval task. The intrinsic challenges in place recognition exist that the confusing objects such as cars and trees frequently occur in the complex urban scene, and buildings with repetitive structures may cause over-counting and the burstiness problem degrading the image representations. To address these problems, we present an Attention-based Pyramid Aggregation Network (APANet), which is trained in an end-to-end manner for place recognition. One main component of APANet, the spatial pyramid pooling, can effectively encode the multi-size buildings containing geo-information. The other one, the attention block, is adopted as a region evaluator for suppressing the confusing regional features while highlighting the discriminative ones. When testing, we further propose a simple yet effective PCA power whitening strategy, which significantly improves the widely used PCA whitening by reasonably limiting the impact of over-counting. Experimental evaluations demonstrate that the proposed APANet outperforms the state-of-the-art methods on two place recognition benchmarks, and generalizes well on standard image retrieval datasets.

## CCS CONCEPTS

• **Computing methodologies** → **Image representations**;

## KEYWORDS

Place recognition; Content-based image retrieval; Convolutional neural network; Attention mechanism

## ACM Reference Format:

Yingying Zhu, Jiong Wang, Lingxi Xie, and Liang Zheng. 2018. Attention-based Pyramid Aggregation Network for Visual Place Recognition. In *2018 ACM Multimedia Conference (MM '18)*, October 22–26, 2018, Seoul, Republic of

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

*MM '18*, October 22–26, 2018, Seoul, Republic of Korea

© 2018 Association for Computing Machinery.

ACM ISBN 978-1-4503-5665-7/18/10...\$15.00

<https://doi.org/10.1145/3240508.3240525>

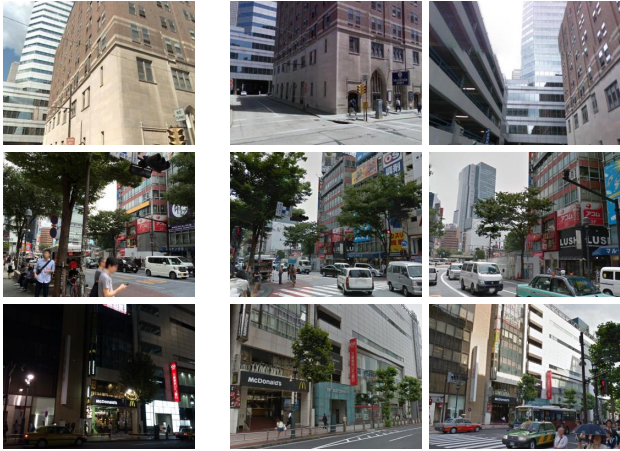
Korea. ACM, New York, NY, USA, 9 pages. <https://doi.org/10.1145/3240508.3240525>

## 1 INTRODUCTION

Visual place recognition has received a considerable level of attention in the community for its wide applications in augmented reality [7, 30], autonomous navigation [14, 29] and 3D reconstruction [1, 10]. Traditionally, visual place recognition has been cast as an image retrieval task at city-scale. Given a query image depicting the scene of a particular location, we aim to find most similar images as location suggestions by querying a large geo-tagged database.

Visual place recognition focuses on the complex urban environment, which contains buildings with repetitive structures [43] and suffers from changes of illumination conditions, seasons or structural modifications over time [42]. Some representative retrieval examples are shown in Figure 1. An inherent problem exists that the trees and cars frequently occur in the urban environment to cause confusion. The buildings (which are geo-informative) with repetitive structures may also cause the problem of over-counting and burstiness [20], *i.e.*, similar descriptors appear such much times in an image as to degrade the image representation. Moreover, partial occlusions and the changes of viewpoint make place recognition a challenging task.

Conventional image retrieval techniques such as bag-of-visual-words (BOW) representation [40] based on local invariant features [28] and vector of locally aggregated descriptors (VLAD) [21] had been adopted to build accurate place recognition systems [4, 25, 42, 43]. An innovative work [2] learned discriminative convolutional neural network (CNN) representations on the Street View training datasets and proposed the NetVLAD layer that implements VLAD by series of differentiable operations. The NetVLAD representations yield competitive results on place recognition and image retrieval datasets. To suppress the confusing elements, Kim *et al.* [24] extend NetVLAD by learning the contextual weights for the local CNN features. But these contextual weights may act as the burstiness frequencies, thus being restricted by intra-normalization [3] in the NetVLAD layer. [24] reports similar performances to NetVLAD on Pitts250k-test dataset [43] and Tokyo24/7 daytime subset [42]. Therefore [24] has not effectively alleviated the influence of confusing objects in the urban environment. In addition, the NetVLAD layer produces high-dimensional representations (*e.g.*,



**Figure 1: Retrieval examples from place recognition datasets. Left: query images. Right: retrieved database images of the same places.**

$64 \times 512$  for VGGNet [39]) and requires a large PCA whitening matrix for dimensionality reduction.

Toward accurate and fast place recognition, we propose an attention-based pyramid aggregation network (APANet), an end-to-end trainable model, which consists of three additional blocks on the base CNN architectures: the spatial pyramid pooling block, attention block, and sum pooling block. As shown in Figure 2, the spatial pyramid pooling is employed to aggregate the multi-size regions on the CNN feature maps. We found this region-based pooling method is effective for encoding the multi-size buildings with repetitive architecture than the global pooling methods [5, 38]. The attention block helps to weight the regional features according to the distinctiveness and then sum pooling produces a compact global descriptor.

In a nutshell, the attention mechanism can be regarded as a strategy that selectively focuses on the informative visual elements, similar to the human perception process. Recently it was introduced to deep neural networks as a powerful addition in a series of vision tasks [18, 44, 48]. Inspired by these successful applications, we incorporate the attention block to suppress the confusing elements in the urban environment. Specifically, we employ two kinds of attention blocks, a single attention block and a cascaded attention block with content prior, to weight the regional features by their distinctiveness. The discriminative regional features will be assigned higher weights than the confusing ones, so that they contribute more to the global descriptor.

In testing stage, we found the widely used PCA whitening approach address the over-counting in an extreme way. Therefore we develop a PCA power whitening strategy which reasonably address the problem of over-counting to get a maximum level of improvement on the retrieval performance. Different with recent works [6, 35] which learn discriminative dimension reduction and whitening using labeled image pairs, PCA power whitening is in a fully unsupervised way and consistently improves PCA whitening without extra computations.

Our APANet is trained in an end-to-end manner on the Street View training datasets [2] targeting for place recognition. On two place recognition benchmark datasets, APANet representations outperform NetVLAD on the same representation dimensionality especially when the dimensionality is low. On the standard image retrieval datasets, APANet surpasses NetVLAD with 8-times more compact representations.

## 2 RELATED WORK

### 2.1 Visual Place Recognition.

Visual Place Recognition in the urban environment is a challenging task due to the frequently occurred confusing objects, the repetitive structures, changes of viewpoint and illumination condition. Based on the bag-of-features model, some previous works [4, 25] focused on discovering distinctive and confusing local descriptors, thereby exploiting selecting or weighting strategies. Torii *et al.* [43] explicitly detected the repetitive image structures and developed an efficient representation of the repeated structures for place recognition in the urban environments. In [42], view synthesis is combined with densely sampled VLAD descriptors to enable robust recognition against variations in viewpoint and illumination.

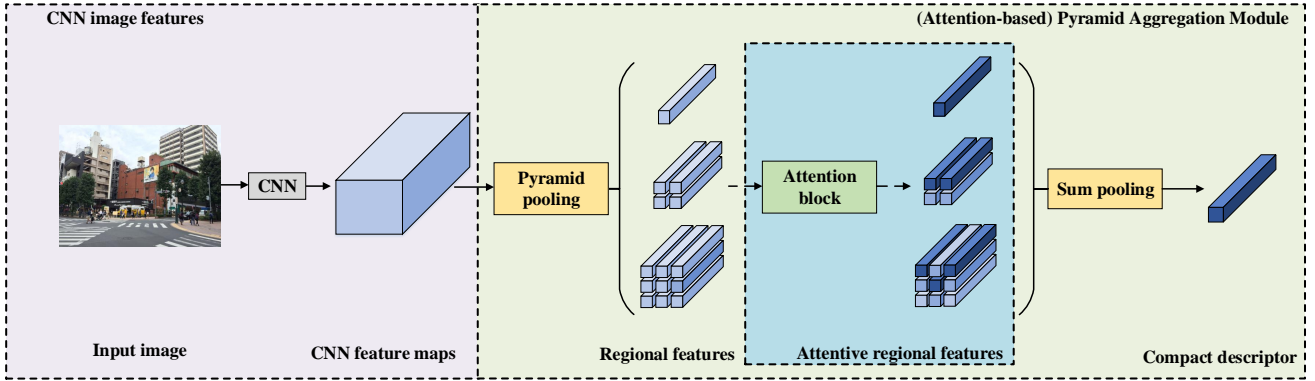
Arandjelović *et al.* [2] performed learning for place recognition and proposed the NetVLAD representations, which significantly outperform the local-feature-based representations on place recognition benchmarks. However, NetVLAD representations may be degraded by the confusing objects and need a large PCA whitening matrix for dimensionality reduction. In contrast, the proposed APANet produces compact representations and the built-in attention blocks can effectively suppress the confusing objects.

### 2.2 CNN-based Image Retrieval.

The seminal work [26] has discussed the feasibility of CNN features for image retrieval. [6, 37] extracted CNN activations from the fully-connected (FC) layer as global descriptors and got preliminary results for image retrieval. However, extracting a single feature vector from the FC layer requires a fixed input image size. Subsequent works exploited the global pooling [5, 38] or region-based pooling methods [41] on activations of the intermediate convolutional layer. Among these works, Tolia *et al.* [41] propose the R-MAC descriptor that aggregates the regional features generated by three scale rigid grids. The regional features are  $L_2$ -normalized, then PCA whitened and  $L_2$ -normalized again before sum-aggregation. In combination with re-ranking and query expansion [8, 9], R-MAC reports competitive performance.

The above-mentioned works adopt pre-trained CNNs while recent state-of-the-art works focus on fine-tuning the pre-trained CNNs on the domain-specific datasets [2, 6, 35]. A pioneering work [6] fine-tuned CNN models on a collected Landmark dataset with cross-entropy loss, which improves the retrieval performance a lot. More recent works [2, 12, 51] reveal the effectiveness of the triplet ranking loss for CNNs fine-tuning in image retrieval and place recognition task.

As a region-based aggregation method, our APANet is close to the works of R-MAC [41] and Deep Image Retrieval (DIR) [12] which learns R-MAC representation in a cleaned landmark dataset. We consider DIR as the baseline method and our APANet differs



**Figure 2: Overview of the proposed APANet. For visualization, the pyramid pooling above adopts the spatial grids with scale (1,2,3). In our practice, we adopt a finer scales, e.g., scale (2,4,6,8). Same as conventional practice, the final compact descriptor are  $L_2$ -normalized by default.**

from it in the following aspects. First, we do not normalize or whiten the regional features before sum-aggregation. In our practice, these two operations perform well for the pre-trained CNNs but are unfavorable for CNN representation learning. Second, we have a finer scale choice for the spatial grids in APANet and increase the scale number to four, which helps the spatial grids prone to align with all the buildings.

The attention block in our APANet helps to weight the regional features with regard to the distinctiveness and there are similar practices in CNN-based image retrieval tasks [16, 22, 23, 32]. Kalantidis *et al.* [23] improves sum pooling by exploiting the spatial weight and channel weight for each location on the feature maps. In [22, 31], novel saliency priors are introduced for aggregating local CNN features. However, to our knowledge, the effectiveness of these weighting strategies have not been demonstrated for fine-tuned CNNs, or for evaluating the regional features. In contrast to these methods, the attention block is involved in the CNN architecture and optimized in the representation learning process, which enables an optimal weighting mechanism for aggregating regional features.

### 3 ATTENTION-BASED PYRAMID AGGREGATION NETWORK

In this section, we describe the proposed attention-based pyramid aggregation network. We first introduce our pyramid aggregation module and show the modifications on pyramid pooling block for place recognition (Section 3.1). Then we depict the attention blocks adopted for weighting the regional features (Section 3.2). Section 3.3 presents the training objective.

#### 3.1 Pyramid Aggregation Module

Our pyramid aggregation (PA) module is composed of spatial pyramid pooling and sum pooling operations. Spatial pyramid pooling [13, 27] was first introduced to CNN by [15] to meet the fixed-length requirement of the fully-connected layer for visual recognition. Recently, it benefits scene parsing [50] and saliency detection [46] tasks by encoding the contextual information. In PA module, spatial

pyramid pooling acts as a region-based pooling method that helps encode the multi-size buildings in an image. To get a compact and discriminative descriptor, the PA module has modifications on the conventional pyramid pooling [15, 50] in two aspects: (i) overlapping max pooling is utilized in the spatial grids so that they can better align with all the buildings, and (ii) all the regional features are sum-aggregated into a compact global descriptor rather than being concatenated.

As shown in Fig. 2, given the feature maps from the CNN’s last convolutional layer, we leverage pyramid pooling on them to aggregate the multi-scale regions. Considering the convolutional feature maps of size  $W \times H \times D$ , where  $W \times H$  is the spatial resolution and  $D$  is the number of channels, the pyramid pooling has pooling window size in proportion to the size of feature maps. For a spatial grid with scale  $n$ , the output is fixed to  $n \times n \times D$  and the size of pooling window is  $\lceil [2 \times W / (n + 1)] \rceil, \lceil [2 \times H / (n + 1)] \rceil$ , stride is  $\lceil [W / (n + 1)] \rceil, \lceil [H / (n + 1)] \rceil$  to enable approximately 50% overlapping on each side, where  $\lceil \cdot \rceil$  is the ceiling operation. Then the regional features are arranged by scale as follow:

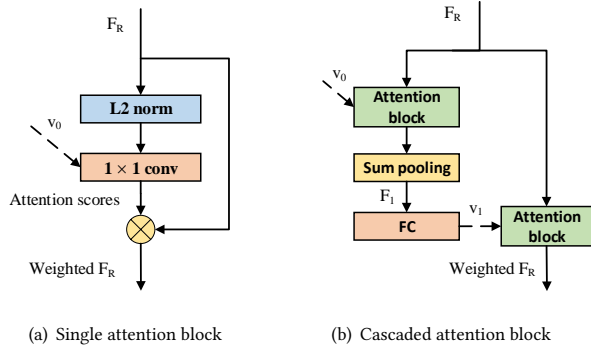
$$\mathbf{F}_R = \left[ \mathbf{f}_{r,1} \dots \mathbf{f}_{r,s_1^2} \dots \mathbf{f}_{r,N} \right], \quad N = \sum_{j=1}^n s_j^2, \quad (1)$$

where  $\mathbf{f}_{r,i}$  is the  $i^{th}$  regional feature with a size of  $1 * 1 * D$ ,  $N$  the total number of regions,  $S_j$  the size of  $j^{th}$  scale, and there are  $n$  scales in total. Thus we obtain a regional feature set  $\mathbf{F}_R$  of size  $1 \times N \times D$  and feed it to the sum pooling block to get the  $1 \times 1 \times D$  global descriptor  $\mathbf{F}_0$ ,

$$\mathbf{F}_0 = \sum_{i=1}^N \mathbf{f}_{r,i}. \quad (2)$$

#### 3.2 Attention Block

In the pyramid aggregation module, all the regional features are sum-aggregated into a global descriptor. A critical problem exists that not all the regional features describe the regions of interest. The receptive fields of some regional features may be centered on



**Figure 3: Schema of single attention block and cascaded attention block.**  $\otimes$  denotes element-wise multiplication operation.

the background or confusing objects like cars and trees in the urban environment. When sum-aggregation, these confusing regional features should be assigned lower weights to suppress their contributions. More precisely, we aim to assign each regional feature a score according to the distinctiveness. Inspired by the attention mechanism popular in fine-grained recognition [47, 48] and video face recognition [49] tasks, we adopt the attention block to evaluate the regional features in PA module.

**3.2.1 Single Attention Block.** Our single attention block utilizes an  $1 \times 1 \times D$  convolutional layer on the regional features  $F_R$  for evaluating their distinctiveness (Fig. 3(a)). The  $D$ -dimensional evaluation vector  $v_0$  (parameters of the  $1 \times 1$  convolutional layer) inner-products all the regional features and produces attention scores, which will multiply the corresponding regional features and get weighted regional features. In this manner, the global descriptor  $F_1$  after sum pooling can be expressed as following:

$$F_1 = \sum_{i=1}^N a_i f_{r,i}, \quad a_i = (v_0 * f_{r,i}), \quad (3)$$

where  $a_i$  is the attention score corresponding to the  $i^{th}$  regional feature,  $f_{r,i}$  is the  $i^{th}$  regional feature after  $L_2$ -normalization, and  $*$  denotes the inner product operation. We normalize the regional features when computing the attention scores because this helps the attention scores be within a reasonable range, and we do not need to adopt an activation function on them.

**3.2.2 Cascaded Attention Block.** A single attention block works as an evaluator for the regional features with regard to their distinctiveness. What if this evaluator has a content prior from the whole image? That is, the evaluator looks through the whole image first and then evaluates each region. We envision that this allows the evaluator to produce more reasonable scores for sum-aggregation. To this end, we borrow the idea of the cascaded attention block which is used for video face recognition [49] and adjust it for weighting the regional features. The cascaded attention block incorporates two attention blocks as shown in Fig. 3(b). The first attention block is the same as the scheme of Fig. 3(a), that produces attention scores for the regional features. Having the global descriptor  $F_1$  in Eq. (3)

after sum pooling, we impose a linear transformation on it by a fully-connected layer followed by hyperbolic tangent nonlinear activation function (tanh). The  $D$ -dimensional output  $v_1$  is used as a new evaluation vector for the second attention block,

$$v_1 = \tanh(WF_1), \quad (4)$$

where  $W$  and  $b$  are the parameters of the fully-connected layer. Compared to  $v_0$  which is randomly initialized in the single attention block, the new evaluation vector  $v_1$  has a content prior from the global image descriptor  $F_1$ .

By incorporating the attention block to the PA module and optimizing it with representation learning process, it can automatically learn to focus on the most discriminative regional features for place recognition. The effectiveness of the proposed attention block will be shown in Section 5.2.

### 3.3 Learning Discriminative Representation with Triplet Ranking Loss

The PA module and the attention block are composed of differentiable conventional CNN operations, thus the proposed APANet is an end-to-end trainable model. Following NetVLAD, we fine-tune APANet with the weakly supervised triplet ranking loss on the Street View training datasets. Triplet loss has shown competitive effectiveness in deep metric learning [17, 45], face identification [36] and image retrieval tasks [2, 12]. Learning to rank the positive and negative images in the triplets enables the network to produce discriminative descriptors. Detailed descriptions about the adopted weakly supervised triplet ranking loss and strategy for mining the training tuples can be found in [2].

## 4 PCA POWER WHITENING

**PCA Whitening.** Whitening is an effective post-processing approach in image retrieval. It helps to solve the problem of over-counting and co-occurrences [19], thereby improving the retrieval accuracy for a series of works [2, 5, 41]. The whitening parameters are usually learned by PCA on an independent dataset. Let us consider a PCA rotation on the descriptors  $X$ ,

$$X_R = P^T X, \quad (5)$$

where  $P$  is the  $D \times D$  PCA rotation matrix,  $X$  is  $L_2$ -normalized and optionally zero-centered. After PCA rotation, the first few dimensions of  $X_R$  preserve most energy and have larger variances than the later dimensions as shown in Fig. 4(a). Meanwhile, the over-counting in the representations (where  $P$  was learned) is captured in  $P$  and mostly influences the first few dimensions of  $X_R$ <sup>1</sup>. Whitening operation (Equation 6) address over-counting problem by balancing the energy (i.e. variance) of each dimension in  $X_R$ . As illustrated in Fig. 4(b), whitening scales each dimension of  $X_R$  to unit variance, which has a pretty enhancement on the performance.

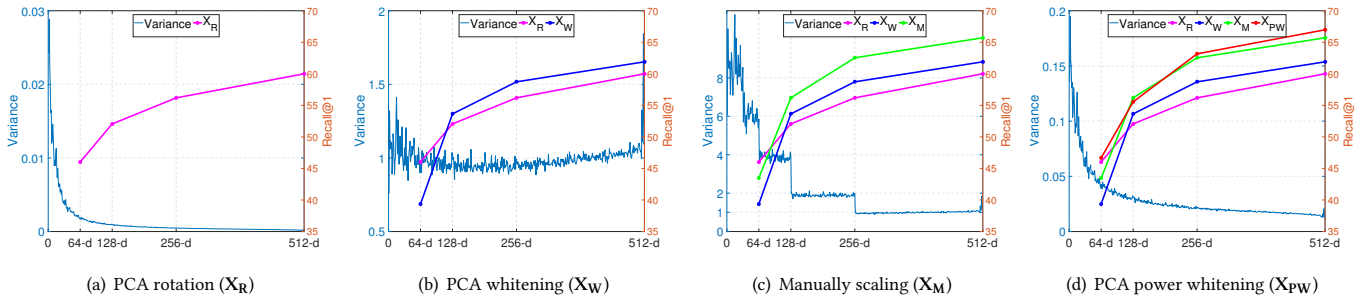
$$X_W = \text{diag}(\lambda_1^{-\frac{1}{2}}, \dots, \lambda_D^{-\frac{1}{2}}) P^T X, \quad (6)$$

where  $\lambda_i$  is the eigenvalue associated with  $i^{th}$  eigenvector in  $P$ .

**Our observation.** However, we argue that PCA whitening may excessively penalize the problem of over-counting. In fact, it is beneficial that the variance of the former dimensions are somehow

<sup>1</sup>A detailed description can be seen in Section 4 of [19].





**Figure 4: Variance versus place recognition accuracy.** For visualization, variances are calculated from the 512-D APANet representations on Tokyo 24/7 database images before normalization while the recalls are from normalized representations.

preserved, so that a balance is achieved between reducing over-counting and preserving the energy distribution of the features. On the basis of  $X_W$ , we manually increase the variances of the first 256 dimensions with different multiples and the performance improvement over  $X_W$  can be seen in Fig. 4(c).

**PCA power whitening.** Motivated by the observation above, we propose a PCA power whitening (PCA-pw) strategy, which provides more reasonable variance scaling based on the eigenvalue  $\lambda_i$ ,

$$X_{PW} = \text{diag}(\lambda_1^{-\frac{1}{2} \times \alpha}, \dots, \lambda_D^{-\frac{1}{2} \times \alpha}) P^T X, \quad (7)$$

where  $0 \leq \alpha \leq 1$  is the parameter of scaling. Usually 0.5 is a reasonable value for  $\alpha$  and we choose it by default for PCA-pw. The performance of PCA-pw is presented in Fig 4(d). We can clearly observe that PCA-pw enables largest performance improvement compared to the all the methods above.

PCA rotation and PCA whitening can be seen as special cases of PCA-pw where  $\alpha = 0$  and 1, respectively. So the PCA-pw can reasonably limit the impact of over-counting by setting a proper value of  $\alpha$ , and providing a maximum level of improvement on the retrieval performance. This argument will be further supported by more experiment results in section 5.3. In our practice, PCA-pw improves a series of retrieval baselines [2, 5, 41] which use PCA whitening.

## 5 EXPERIMENT

In this section, we present the place recognition datasets, analyze the scale choice for pyramid pooling block and demonstrate effectiveness of the attention block. Then we apply PCA power whitening on four representative aggregation methods. Finally, we compare APANet to the state-of-the-art and show its generalization ability on the standard image retrieval datasets.

### 5.1 Datasets and Implementation Details

We evaluate the proposed APANet on two place recognition datasets: **Pitts250k-test** the [43] and **Tokyo 24/7** [42] datasets. Pitts250k contains 254k perspective images generated from 10.6k Google Street View panoramas in Pittsburgh area. Pitts250k-test is a subset of Pitts250k and has around 83k database images, 8k query images. Tokyo 24/7 is a challenging dataset that contains 76k database images and 315 query images captured by different mobile phones cameras at daytime, sunset and night. The Street View

**Table 1: Performance of R-MAC and PANet with different scale choices and SNW operations. PANet (1234) has four scales spatial grids and 30 regional features (1, 2×2, 3×3, 4×4). The best results are highlighted in bold. All these methods are based on AlexNet.**

Method	Regions	Pitts250k-test		
		R@1	R@5	R@10
R-MAC w/o SNW	20	68.35	82.93	87.63
PANet (1234)	30	68.86	83.47	87.57
PANet (2468)	120	<b>69.69</b>	<b>83.77</b>	87.85
PANet (2345678)	203	69.24	83.19	87.23
R-MAC [12]	20	68.24	83.70	<b>87.97</b>
PANet (1234) + SNW	30	67.27	82.56	86.70
PANet (2468) + SNW	120	63.60	79.78	84.65
PANet (2345678) + SNW	203	67.21	82.11	86.35

training datasets [2] consist of Pitts30k-train and Tokyo Time Machine (TokyoTM) dataset. We choose the Pitts30k-train or TokyoTM dataset for fine-tuning according to the testing dataset.

**Evaluation metric.** For these two evaluation datasets, we follow the standard evaluation protocol in [2, 42, 43]. The performance is measured by the percentage of correctly recognized queries at given top  $N$  candidates (Recall@ $N$ ). A query image is deemed to be correctly recognized if at least one of the top  $N$  candidate database images are within 25 meters from the ground truth position.

**Implementation details.** The pre-trained AlexNet [26] and VGG-16 [39] architectures are adopted as the base CNN architectures for fine-tuning and both are cropped at the last convolutional layer before Relu. Besides the attention-based pyramid aggregation module, the R-MAC and sum pooling are also adopted as the aggregation layer for fine-tuning. For all these methods, we use margin  $m = 0.1$ , batch size of 4 tuples, SGD with initial learning rate  $l_0$  of 0.001 for Pitts30k-train and 0.0005 for the TokyoTM dataset, and an exponential decay  $l_0 \exp(-0.1(i-1))$  over epoch  $i$ , momentum 0.9, weight decay 0.001. We use Xavier initialization [11] for the attention blocks, whose learning rate is ten times of the formal convolutional layers. When testing, the whitening parameters are learned from 10k images randomly sampled from the Pitts30k-train or TokyoTM dataset, the same as NetVLAD. For fair competition

Table 2: Comparison of two aggregation methods when the single or cascaded attention block is integrated. All these methods are based on AlexNet. The best results of each method are highlighted in bold.

Method	Single	Cascaded	Pitts250k-test			Tokyo 24/7		
			Recall@1	Recall@5	Recall@10	Recall@1	Recall@5	Recall@10
Sum pooling	√	√	58.53	75.56	82.08	28.57	42.22	53.02
			60.76	77.23	82.62	28.25	<b>46.35</b>	53.33
			<b>63.84</b>	<b>78.95</b>	<b>84.12</b>	<b>29.84</b>	46.03	<b>54.60</b>
PANet			69.69	83.77	87.85	33.65	48.57	53.02
APANet	√	√	71.20	85.75	89.41	34.92	49.21	53.65
			<b>72.69</b>	<b>86.38</b>	<b>89.92</b>	<b>38.41</b>	<b>53.97</b>	<b>61.27</b>

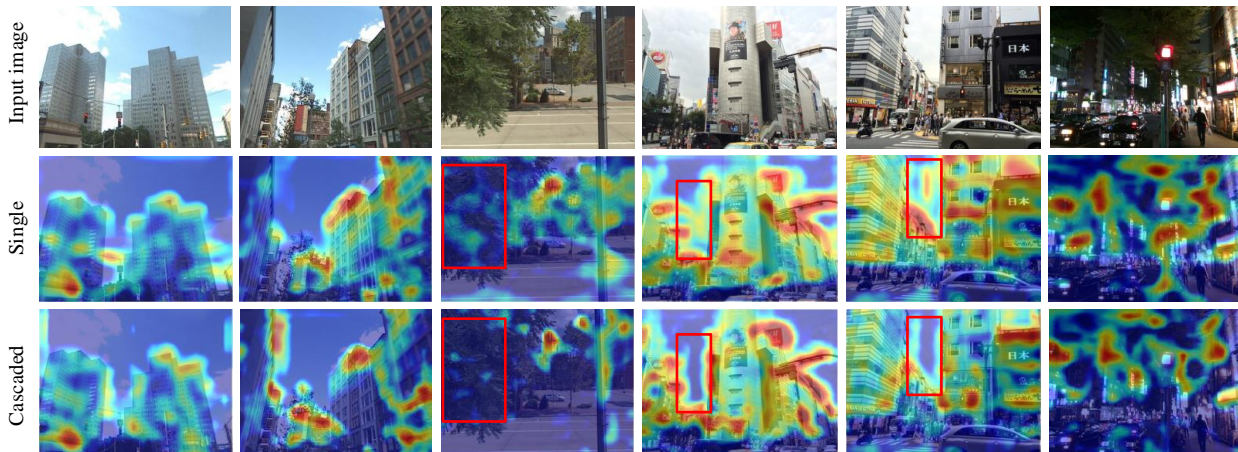


Figure 5: Visualization of attention score maps in heat maps. (top) input images. (middle) attention scores from a single attention block defined in Fig. 3(a). (bottom) attention scores from the cascaded attention block defined in Fig. 3(b).

with NetVLAD, we do not perform data augmentation or three-clip testing as [24] did.

## 5.2 Evaluation of APANet

**5.2.1 Scale Choice and Baseline.** We consider R-MAC as the baseline method and fine-tune the base architecture with R-MAC on the Street View training datasets. The network configuration is the same as DIR [12]. Our PA module differs from R-MAC by adopting a different region choice and removing the shift, normalization and whitening (SNW) operations on the regional features.

We first analyze the scale choice for our pyramid pooling block. R-MAC has three scales of rigid grids which define around 20 regions ( $1 \times 2$ ,  $2 \times 3$ ,  $3 \times 4$ ). The grid scale of R-MAC is too coarse because the receptive field of each grid cell covers the whole image, which is ineffective for encoding the local cues in an image for complex scenes in place recognition. Hence, we adopt a finer scale choice in the proposed PA module (*i.e.* PANet) and increase the scale number to four. The upper part of Table 1 presents the comparisons of different region choices. We observe that “PANet (2468)” exceeds “PANet (1234)” and “R-MAC w/o SNW” by adopting a finer scale choice and more region numbers. This is because the likelihood that the regions of interest are well-aligned increases as the number of regions increases. However, increasing the number of regions may also incur more confusing regions to corrupt the image similarity

measurement. It can be proven that “PANet (2345678)” with the largest number of scales and regions does not perform best.

Then we discuss the SNW operations in R-MAC and PANet. As shown in Table 1, R-MAC performs similarly when attaching or removing SNW operations. But the performance of PANet with SNW operations drops significantly when the scales get finer. The interpretation is that the number of confusing regional features increases when the scale gets finer. These confusing regional features usually carry lower norms than the discriminative ones while SNW operations may highlight their contributions. Thus in the following experiments, we use four grid scales (2, 4, 6, 8) for the pyramid pooling block by default and do not adopt SNW operations.

**5.2.2 Effect of Attention Block.** Note that attention blocks can also weight the local CNN features for sum pooling method. We evaluate the attention blocks by employing them on the sum pooling and our PA module (namely APANet). Training is conducted on the Street View training datasets and the results are displayed in Table 2. We observe that the attention blocks improve all these two aggregation methods on two datasets and the cascaded attention block always works better than the single one. In addition, the PA module has significantly better performances than sum pooling, which also indicates the effectiveness of PA module.

For visualization, attention scores of sum pooling are presented in Fig. 5. These two attention blocks really work as we expected, *i.e.*,

**Table 3: Comparison of PCA whitening and PCA power whitening (PCA-pw). All the results are from the 512-D representations based on VGG-16 architecture.**

Method	Whitening	Pitts250k-test			Tokyo 24/7		
		Recall@1	Recall@5	Recall@10	Recall@1	Recall@5	Recall@10
Mac	W/o whitening	77.01	88.73	91.97	38.41	52.70	62.22
	PCA whitening	73.21	86.03	89.77	25.40	40.63	45.40
	PCA-pw	<b>79.19</b>	<b>90.12</b>	<b>93.09</b>	35.56	52.06	60.95
	PCA-pw ( $\alpha = 0.1$ )	78.25	89.44	92.26	<b>38.73</b>	<b>53.97</b>	<b>62.54</b>
Sum pooling	PCA whitening	74.13	86.44	90.18	44.76	60.95	70.16
	PCA-pw	<b>75.63</b>	<b>88.01</b>	<b>91.75</b>	<b>52.70</b>	<b>67.30</b>	<b>73.02</b>
NetVLAD [2]	PCA whitening	80.66	90.88	93.06	<b>60.00</b>	73.65	79.05
	PCA-pw	<b>81.95</b>	<b>91.65</b>	<b>93.76</b>	58.73	<b>74.60</b>	<b>80.32</b>
APANet	PCA whitening	82.32	90.92	93.79	61.90	77.78	80.95
	PCA-pw	<b>83.65</b>	<b>92.56</b>	<b>94.70</b>	<b>66.98</b>	<b>80.95</b>	<b>83.81</b>

focusing on the architectures and assigning lower attention scores to confusing objects such as pedestrians, cars and trees. Fig. 5 also suggests that cascaded attention block has more localized attention score maps than the single one, thus paying more attention to the most discriminative regional features and resulting in more discriminative image representations. This observation can be viewed at column 3-5. For example, the buildings are severely occluded by trees at column 3. In this case, the cascaded attention block can still focus on the buildings while in comparison the single attention block is less concentrated. At columns 4 - 5, the background between two buildings is usually assigned lower scores in the cascaded attention score maps while it is considerably activated in single attention score maps. In rest of the paper, we use cascaded attention block for our APANet unless otherwise specified.

### 5.3 PCA Power Whitening

To assess the effectiveness and universality of the proposed PCA power whitening (PCA-pw), we compare it with PCA whitening on four representative aggregation methods, *i.e.* global max pooling (Mac), sum pooling, NetVLAD and our APANet. We learn the image representations with these aggregation methods on the Street View training datasets and present the results in Table 3. Several things can be observed. First, PCA-pw usually performs better than PCA whitening, especially on the Tokyo 24/7 dataset where the over-counting problem from the buildings is not so serious than Pitts250k-test. Second, for the Mac representations, the problem of over-counting is not a big deal. PCA Whitening even decreases the performances on both datasets because it excessively penalizes the over-counting. By alleviating these, PCA-pw improves the performance of Mac representations on Pitts250k-test dataset, but still decreases on Tokyo 24/7. We find there is a small enhancement on Tokyo 24/7 when setting the scaling factor  $\alpha$  to 0.1, which reflects the degree of over-counting that Mac representations suffer on two datasets. Third, APANet representations perform best on both datasets, regardless of which whitening strategy is adopted.

In summary, because of the aggregation method, dataset characteristics and the local CNN features themselves, the CNN-based image representations may suffer from over-counting problem more or less. The over-counting is solved by PCA Whitening in an extreme way while the proposed PCA-pw can better address it by

setting a reasonable scaling factor  $\alpha$ , thereby providing consistently performance improvement for image retrieval.

### 5.4 Comparison with State-of-the-Art

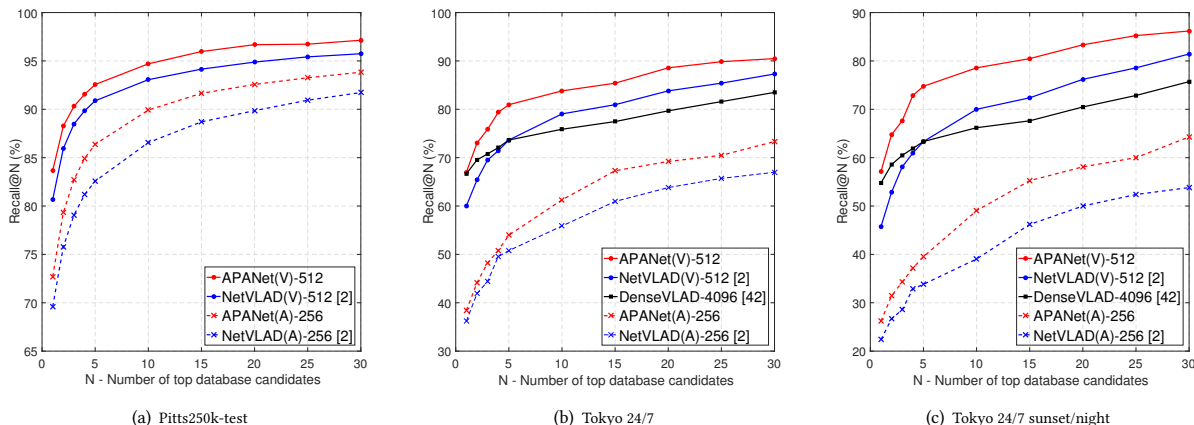
The Dense-VLAD [12] combining view synthesis with densely sampled VLAD descriptors enables recognition across the variations in viewpoint and illumination condition. And the NetVLAD-based deep representations [2, 24] achieve the state-of-the-art performance in place recognition datasets. We compare APANet with these methods and present the result in Fig. 6<sup>2</sup>. It can be seen that APANet consistently outperforms NetVLAD using the AlexNet and VGG-16 architecture on all the datasets. For VGG-16 architecture, the Recall@1 accuracies of APANet exceed NetVLAD with margins of 2.99% and 6.98% on Pitts250k-test and Tokyo 24/7 dataset, respectively. Even if we didn't perform data augmentation on the lighting conditions, the gap between APANet(V) and the Dense-VLAD is even more pronounced on the challenging Tokyo 24/7 sunset/night subset, which demonstrates that the proposed APANet is robust to changes of illumination and viewpoint.

**Dimensionality reduction with PCA power whitening.** To further assess the performance of APANet, we present the performance of APANet and NetVLAD on varying representation dimensionalities in Fig. 7. We observe that PCA-pw consistently outperforms PCA whitening for APANet representations from low to high dimensionality. Meanwhile PCA-pw improves NetVLAD on the Pitts250k-test dataset but the improvements are not so stable on the Tokyo 24/7. We speculate there may be a more suitable scaling factor  $\alpha$  for NetVLAD representations on Tokyo 24/7 dataset. Moreover, compared to the NetVLAD, the recall@5 curves of APANet decrease gracefully with dimensionality reduction. For similar performance, APANet representations are usually two-times more compact than NetVLAD. This phenomenon is more pronounced on the challenging Tokyo 24/7 dataset.

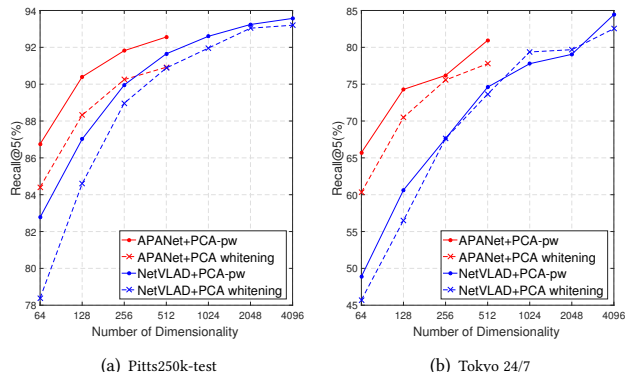
### 5.5 Instance Image Retrieval

To evaluate the generalization ability of APANet, we deploy the APANet model (trained on Pitts30k-train dataset) on two standard

<sup>2</sup>We do not include the curves of [24] in the figure because we can not get the recall curves or the trained models from the authors, and the performance of [24] is similar to NetVLAD on Pitts250k-test dataset.



**Figure 6: Comparison of recalls with previous state-of-the-arts. The base CNN architecture is denoted in brackets: (V) for VGG-16 and (A) for AlexNet model. The dimensionality is followed.**



**Figure 7: Comparison of place recognition accuracy and dimensionality.**

image retrieval benchmarks, the Oxford5k [33] and Paris6k [34] datasets. We have thorough comparisons on dimensionality with NetVLAD. The results are displayed in Table 4. All the results are based on the single scale image representations and no spatial re-ranking or query expansion is adopted. We observe that when learning the whitening parameters from Pitts30k-train dataset as NetVLAD does, APANet even outperforms the 4096-D NetVLAD representations on both two datasets by 512-D representations, and it still performs well with extremely short codes. Further, when learning the whitening parameters from the Oxford5k or Paris6k dataset representations as conventional practices do, APANet gets consistent performance improvements from high to low dimensionality.

## 6 CONCLUSIONS

In this paper, we propose an APANet which is well-designed to overcome the challenges in place recognition task. Experiments demonstrate that APANet representations are robust to changes of viewpoint and illumination and outperform NetVLAD using

**Table 4: Comparison with NetVLAD on image retrieval datasets. The accuracy is measured by mean average precision (mAP) and these methods are based on VGG-16 architecture. † denotes that the results at the column are from representations whitened on Pitts30k-train dataset and ‡ denotes Oxford5k or Paris6k.**

Dim	Oxford5k		Paris6k	
	NetVLAD[2]	APANet	NetVLAD[2]	APANet
4096	71.6 <sup>†</sup>	-	79.7 <sup>†</sup>	-
2048	70.8	-	78.3	-
1024	69.2	-	76.5	-
512	67.6	75.1 <sup>†</sup> 77.9 <sup>‡</sup>	74.9	80.2 <sup>†</sup> 83.5 <sup>‡</sup>
256	63.5	72.8 75.6	73.5	76.9 81.7
128	61.4	67.3 71.7	69.5	74.8 78.7
64	51.1	58.5 63.9	63.0	70.7 73.0
32	42.6	46.4 48.7	54.4	62.5 63.7
16	29.9	31.7 33.4	44.9	48.3 52.4

the same or even lower dimensional representations. Meanwhile, APANet emerges powerful generalization ability on standard image retrieval datasets. In addition, the proposed PCA power whitening strategy consistently improves performance for APANet and is applicable for other retrieval tasks as well. In our future works, we will improve our APANet for instance image retrieval task.

## ACKNOWLEDGMENTS

This work was supported by: (i) National Natural Science Foundation of China (Grant No. 61602314); (ii) Natural Science Foundation of Guangdong Province of China (Grant No. 2016A030313043); (iii) Fundamental Research Project in the Science and Technology Plan of Shenzhen (Grant No. JCYJ20160331114551175). We would also like to thank Relja Arandjelović and Akihiko Torii for providing data, codes, and sharing insights, and Jie Lin for insightful discussions.



## REFERENCES

- [1] Sameer Agarwal, Yasutaka Furukawa, Noah Snavely, Ian Simon, Brian Curless, Steven M Seitz, and Richard Szeliski. 2011. Building rome in a day. *Commun. ACM* 54, 10 (2011), 105–112.
- [2] Relja Arandjelović, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. 2016. NetVLAD: CNN architecture for weakly supervised place recognition. In *CVPR*. 5297–5307.
- [3] Relja Arandjelović and Andrew Zisserman. 2013. All About VLAD. In *CVPR*. 1578–1585.
- [4] Relja Arandjelović and Andrew Zisserman. 2014. DisLocation: Scalable descriptor distinctiveness for location recognition. In *ACCV*. 188–204.
- [5] Artem Babenko and Victor Lempitsky. 2015. Aggregating local deep features for image retrieval. In *ICCV*. 1269–1277.
- [6] Artem Babenko, Anton Slesarev, Alexandr Chigorin, and Victor Lempitsky. 2014. Neural codes for image retrieval. In *ECCV*. 584–599.
- [7] David M Chen, Sam S Tsai, Ramakrishna Vedantham, Radek Grzeszczuk, and Bernd Girod. 2009. Streaming mobile augmented reality on mobile phones. In *ISMAR*. 181–182.
- [8] Ondřej Chum, Andrej Mikulík, Michal Perdoch, and Jiří Matas. 2011. Total recall II: Query expansion revisited. In *CVPR*. 889–896.
- [9] Ondřej Chum, James Philbin, Josef Sivic, Michael Isard, and Andrew Zisserman. 2007. Total recall: Automatic query expansion with a generative feature model for object retrieval. In *ICCV*. 1–8.
- [10] David Crandall, Andrew Owens, Noah Snavely, and Dan Huttenlocher. 2011. Discrete-continuous optimization for large-scale structure from motion. In *CVPR*. 3001–3008.
- [11] Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Aistats*. 249–256.
- [12] Albert Gordo, Jon Almazán, Jerome Revaud, and Diane Larlus. 2016. Deep image retrieval: Learning global representations for image search. In *ECCV*. 241–257.
- [13] Kristen Grauman and Trevor Darrell. 2005. The pyramid match kernel: Discriminative classification with sets of image features. In *ICCV*, Vol. 2. 1458–1465.
- [14] James Hays and Alexei A Efros. 2008. IM2GPS: estimating geographic information from a single image. In *CVPR*. 1–8.
- [15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2014. Spatial pyramid pooling in deep convolutional networks for visual recognition. In *ECCV*. 346–361.
- [16] Tuan Hoang, Thanh-Toan Do, Dang-Khoa Le Tan, and Cheung Ngai-Man. 2017. Selective deep convolutional features for image retrieval. In *ACM MM*. 1600–1608.
- [17] Elad Hoffer and Nir Ailon. 2015. Deep metric learning using triplet network. In *SIABAD*. 84–92.
- [18] Jie Hu, Li Shen, and Gang Sun. 2018. Squeeze-and-Excitation Networks. *CVPR*.
- [19] Hervé Jégou and Ondřej Chum. 2012. Negative evidences and co-occurrences in image retrieval: The benefit of PCA and whitening. *ECCV*, 774–787.
- [20] Hervé Jégou, Matthijs Douze, and Cordelia Schmid. 2009. On the burstiness of visual elements. In *CVPR*. 1169–1176.
- [21] Herve Jegou, Florent Perronnin, Matthijs Douze, Jorge Sánchez, Patrick Perez, and Cordelia Schmid. 2012. Aggregating local image descriptors into compact codes. *TPAMI* 34, 9 (2012), 1704–1716.
- [22] Albert Jiménez, Jose M Alvarez, and Xavier Giró Nieto. 2017. Class-weighted convolutional features for visual instance search. In *BMVC*. 1–12.
- [23] Yannis Kalantidis, Clayton Mellina, and Simon Osindero. 2016. Cross-dimensional weighting for aggregated deep convolutional features. In *ECCV*. 685–701.
- [24] Hyo Jin Kim, Enrique Dunn, and Jan-Michael Frahm. 2017. Learned contextual feature reweighting for image geo-localization. In *CVPR*. 2136–2145.
- [25] Jan Knopp, Josef Sivic, and Tomas Pajdla. 2010. Avoiding confusing features in place recognition. *ECCV*, 748–761.
- [26] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *NIPS*. 1097–1105.
- [27] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. 2006. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, Vol. 2. 2169–2178.
- [28] David G Lowe. 2004. Distinctive image features from scale-invariant keypoints. *IJCV* 60, 2 (2004), 91–110.
- [29] Colin McManus, Winston Churchill, Will Maddern, Alexander D Stewart, and Paul Newman. 2014. Shady dealings: Robust, long-term visual localisation using illumination invariance. In *ICRA*. 901–906.
- [30] Sven Middelberg, Torsten Sattler, Ole Untzelmann, and Leif Kobbelt. 2014. Scalable 6-dof localization on mobile devices. In *ECCV*. 268–283.
- [31] Eva Mohedano, Kevin McGuinness, Xavier Giro-i Nieto, and Noel E O’Connor. 2017. Saliency Weighted Convolutional Features for Instance Search. *arXiv preprint arXiv:1711.10795* (2017).
- [32] Hyeonwoo Noh, Andre Araujo, Jack Sim, Tobias Weyand, and Bohyung Han. 2017. Largescale image retrieval with attentive deep local features. In *ICCV*. 3456–3465.
- [33] James Philbin, Ondřej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman. 2007. Object retrieval with large vocabularies and fast spatial matching. In *CVPR*. 1–8.
- [34] James Philbin, Ondřej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman. 2008. Lost in quantization: Improving particular object retrieval in large scale image databases. In *CVPR*. 1–8.
- [35] Filip Radenović, Giorgos Tolias, and Ondřej Chum. 2016. CNN image retrieval learns from BoW: Unsupervised fine-tuning with hard examples. In *ECCV*. 3–20.
- [36] Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *CVPR*. 815–823.
- [37] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. 2014. CNN features off-the-shelf: an astounding baseline for recognition. In *CVPR Workshops*. 806–813.
- [38] Ali Sharif Razavian, Josephine Sullivan, Stefan Carlsson, and Atsuto Maki. 2016. Visual instance retrieval with deep convolutional networks. *ITE Trans. MTA* 4, 3 (2016), 251–258.
- [39] Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. In *ICLR*.
- [40] Josef Sivic and Andrew Zisserman. 2003. Video Google: A text retrieval approach to object matching in videos. In *ICCV*. 1470–1477.
- [41] Giorgos Tolias, Ronan Sicre, and Hervé Jégou. 2016. Particular object retrieval with integral max-pooling of CNN activations. In *ICLR*.
- [42] Akihiko Torii, Relja Arandjelović, Josef Sivic, Masatoshi Okutomi, and Tomas Pajdla. 2015. 24/7 place recognition by view synthesis. In *CVPR*. 1808–1817.
- [43] Akihiko Torii, Josef Sivic, Tomas Pajdla, and Masatoshi Okutomi. 2013. Visual place recognition with repetitive structures. In *CVPR*. 883–890.
- [44] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. 2017. Residual Attention Network for Image Classification. In *CVPR*. 3156–3164.
- [45] Jiang Wang, Yang Song, Thomas Leung, Chuck Rosenberg, Jingbin Wang, James Philbin, Bo Chen, and Ying Wu. 2014. Learning fine-grained image similarity with deep ranking. In *CVPR*. 1386–1393.
- [46] Tiantian Wang, Ali Borji, Lihe Zhang, Pingping Zhang, and Huchuan Lu. 2017. A Stage-wise Refinement Model for Detecting Salient Objects in Images. In *CVPR*. 4019–4028.
- [47] Tianjun Xiao, Yichong Xu, Kuiyuan Yang, Jiaying Zhang, Yuxin Peng, and Zheng Zhang. 2015. The application of two-level attention models in deep convolutional neural network for fine-grained image classification. In *CVPR*. 842–850.
- [48] Yichao Yan, Bingbing Ni, and Xiaokang Yang. 2017. Fine-Grained Recognition via Attribute-Guided Attentive Feature Aggregation. In *ACM MM*. 1032–1040.
- [49] Jiaolong Yang, Peiran Ren, Dongqing Zhang, Dong Chen, Fang Wen, Hongdong Li, and Gang Hua. 2017. Neural Aggregation Network for Video Face Recognition. In *CVPR*. 2492–2495.
- [50] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. 2017. Pyramid scene parsing network. In *CVPR*. 2881–2890.
- [51] Liang Zheng, Yi Yang, and Qi Tian. 2017. SIFT meets CNN: A decade survey of instance retrieval. *TPAMI* (2017).