24/7 place recognition by view synthesis

Akihiko Torii, Relja Arandjelovic, Josef Sivic, Masatoshi Okutomi, Tomas Pajdla

24/7 place recognition by view synthesis

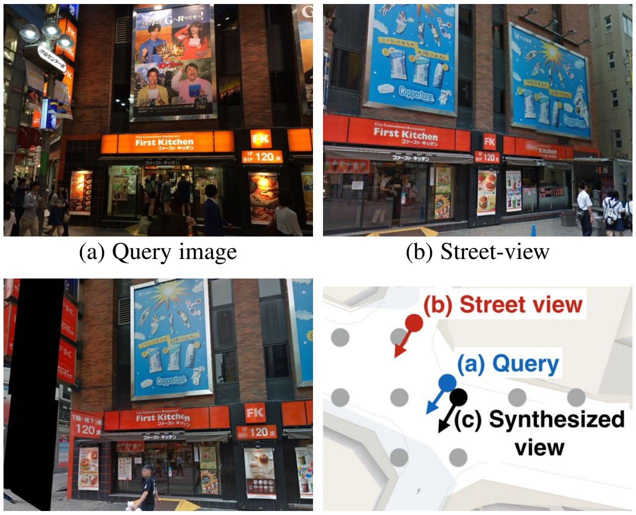
Akihiko Torii, Relja Arandjelović, Josef Sivic, Masatoshi Okutomi, and Tomas Pajdla

Abstract—We address the problem of large-scale visual place recognition for situations where the scene undergoes a major change in appearance, for example, due to illumination (day/night), change of seasons, aging, or structural modifications over time such as buildings being built or destroyed. Such situations represent a major challenge for current large-scale place recognition methods. This work has the following three principal contributions. First, we demonstrate that matching across large changes in the scene appearance becomes much easier when both the query image and the database image depict the scene from approximately the same viewpoint. Second, based on this observation, we develop a new place recognition approach that combines (i) an efficient synthesis of novel views with (ii) a compact indexable image representation. Third, we introduce a new challenging dataset of 1,125 camera-phone query images of Tokyo that contain major changes in illumination (day, sunset, night) as well as structural changes in the scene. We demonstrate that the proposed approach significantly outperforms other large-scale place recognition techniques on this challenging data.

Index Terms—Place Recognition, View Synthesis, Compact Image Descriptor, Image Retrieval.

1 Introduction

Recent years have seen a tremendous progress [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [12], [13], [14], [15], [16] in the large-scale visual place recognition problem [14], [17]. It is now possible to obtain an accurate camera position of a query photograph within an entire city represented by a dataset of 1M images [1], [4], [15] or a reconstructed 3D point cloud [8], [10]. These representations are built on local invariant features such as SIFT [18] so that recognition can proceed across moderate changes in viewpoint, scale or partial occlusion by other objects. Efficiency is achieved by employing inverted file [19], [20] or product quantization [21] indexing techniques. Despite this progress, identifying the same place across major changes in the scene appearance due to illumination (day/night), change of seasons, aging, or structural modifications over time [22], [23], as shown in figure 1, remains a major challenge. Solving this problem would have, however, significant practical implications. Imagine, for example, automatically searching public archives to find all imagery depicting the same place to analyze changes over time for applications in architecture, archaeology and urban planning; or visualize the same place in different illuminations, seasons or backward in time.



(a) Query image (b) Street-view

(c) Synthesized view (d) Locations on the map

Fig. 1. **Matching across major changes in scene appearance is easier for similar viewpoints.** (a) Query image. (b) The original database image cannot be matched to the query due to a major change in scene appearance combined with large change in the viewpoint. (c) Matching a more similar synthesized view is possible. (d) Illustration of locations of (a-c) on the map. The dots and arrows indicate the camera positions and view directions.

In this paper, we demonstrate that matching across large changes in scene appearance is easier when both the query image and the database image depict the scene from approximately the same viewpoint. We implement this idea by synthesizing virtual views on a densely sampled grid on the map. This poses the following three major challenges. First, how can we efficiently synthesize virtual viewpoints for an entire city? Second, how do we deal with the increased database size augmented by the additional synthesized views? Finally, how do we represent the synthetic views in a way that is robust to the large changes in scene appearance?

To address these issues, we, first, develop a view synthesis method that can render virtual views directly from Google street-

- *A. Torii is with the Department of Systems and Control Engineering, the School of Engineering, Tokyo Institute of Technology.*
E-mail: torii@ctrl.titech.ac.jp
- *R. Arandjelović is with the Inria, WILLOW, Departement d'Informatique de l'École Normale Supérieure, ENS/INRIA/CNRS UMR 8548, Paris.*
E-mail: relja.arandjelovic@inria.fr
- *J. Sivic is with the Inria, WILLOW, Departement d'Informatique de l'École Normale Supérieure, ENS/INRIA/CNRS UMR 8548, Paris and the Czech Institute of Informatics, Robotics and Cybernetics, Czech Technical University in Prague.*
E-mail: Josef.Sivic@ens.fr
- *M. Okutomi is with the Department of Systems and Control Engineering, the School of Engineering, Tokyo Institute of Technology.*
E-mail: mxo@ctrl.titech.ac.jp
- *T. Pajdla is with the Center for Machine Perception, Department of Cybernetics, Faculty of Electrical Engineering, Czech Technical University in Prague.*
E-mail: pajdla@cvut.cz

view panoramas and their associated approximate depth maps, not requiring to reconstruct an accurate 3D model of the scene. While the resulting images are often noisy and contain artifacts, we show that this representation is sufficient for the large-scale place recognition task. The key advantage of this approach is that the street-view data is available world-wide opening-up the possibility for a truly planet-scale [24] place recognition. Secondly, to cope with the large amount of synthesized data – as much as nine times more images than in the original street-view – we use the compact VLAD encoding [25], [26] of local image descriptors, which is amenable to efficient compression, storage and indexing. Finally, we represent images using local gradient based descriptors (RootSIFT [18], [27] in our case) *densely sampled* across the image and at multiple scales. We found that this representation is more robust to large changes in appearance due to illumination, aging, *etc*. as it does not rely on repeatable detection of local invariant features, such as the Laplacian of Gaussian [18]. While local invariant features have been successfully used for almost two decades to concisely represent images for matching across viewpoint and scale [28] they are often non-repeatable across non-modeled changes in appearance due to, *e.g*., strong perspective effects or major changes in the scene illumination [29], [30]. Not relying on the local invariant keypoint detection comes at a price of reduced invariance to geometric transformation. However, we have found this is in fact an advantage, rather than a problem, as the resulting representation is more discriminative and thus copes better with the increased rate of false positive images due to the much larger database augmented with synthetic views.

The paper is organized as follows. In section 2 we describe related work on large-scale visual place recognition. In section 3 we investigate the challenges in matching local features across major changes in scene appearance and illustrate benefits of densely sampled descriptors. Our view synthesis method for expanding the database is then described in section 4 and the complete system for place recognition with view synthesis is outlined in section 5. Finally, in section 6 we experimentally demonstrate the benefits of the proposed method on the a new challenging dataset that contains major changes in illumination (day, sunset, night) as well as structural changes in the scenes.

2 Related work

Place recognition with local-invariant features. The large-scale place recognition is often formulated as a variation of image retrieval [19], [31] where the query photograph is localized by matching it to a large database of geo-tagged images such as Google street-view [2], [4], [5], [6], [7], [11], [14], [15], [16]. The 3D structure of the environment can be also reconstructed beforehand and the query is then matched directly to the reconstructed point-cloud [8], [10] rather than individual images. The underlying appearance representation for these methods is based on local invariant features [28], either aggregated into an image-level indexable representation [4], [5], [6], [7], [15], [16], or associated to individual reconstructed 3D points [8], [10], [12]. These methods have shown excellent performance for large-scale matching across moderate changes of scale and viewpoint that are modeled by the local invariant feature detectors. However, matching across non-modeled appearance variations such as major changes in illumination, aging, or season are still a challenge.

We investigate compact representations based on descriptors densely sampled across the image rather than based on local-invariant features. Densely sampled descriptors have been long used for category-level recognition [32], [33], [34], [35] including category-level localization [6], but due to their limited invariance to geometric transformations have been introduced to instance-level recognition only recently [36]. While we build on this work, we show that combining dense representations with virtual view synthesis can be used for large-scale place recognition across significant changes of scene appearance. Densely sampled representations can be also extracted using convolutional neural networks (CNNs) [37], [38]. In contrast to category-level recognition, most of the recent work that uses CNN descriptors for instance-level matching [39], [40], [41], however, did not provide significant improvements over standard RootSIFT descriptors. Significant improvements using CNNs have been obtained only very recently by training the image representation directly for the place recognition task [42], which is complementary to our view synthesis as we demonstrate in section 6.

Virtual views for instance-level matching. Related to our work are also methods that generate some form of virtual data for instance-level matching, but typically they focus on extending the range of recognizable viewpoints [43], [44], [45] or matching across domains [29], [46] and do not consider compact representations for large-scale applications. Irschara *et al.* [43] generate bag-of-visual-word descriptors extracted from existing views for virtual locations on a map to better model scene visibility. Shan *et al.* [44] use 3D structure to synthesize virtual views to match across extreme viewpoint changes for alignment of aerial to ground-level imagery. Wu *et al.* [45] locally rectify images based on the underlying 3D structure to extend the viewpoint invariance of local invariant features (SIFT). Their method has been successfully applied for place recognition [4] but requires either known 3D structure or rectification on the query side. Recently, rendering virtual views has been also explored for cross-domain matching to align paintings to 3D models [29] or to match SIFT descriptors between images and laser-scans [46].

Modelling scene illumination for place recognition. In place recognition, the related work on modeling outdoor illumination has focused on estimating locations and time-stamps from observed illumination effects [47], [48]. In contrast, we focus on recognizing the same scene across changes of illumination. However, if illumination effects could be reliably synthesized [49], [50] the resulting imagery could be used to further expand the image database.

Handling illumination and appearance changes in robot localization. To compensate for day-and-night illumination changes, Maddern *et al.* [51] generate illumination invariant images by conversion from 3-channel (RGB) to a single channel image using peaks of spectral responses of the sensor. However, their method requires a known specification of the sensor hardware for luminance calibration, which is not available in our set-up where query images can come from different sources and devices. To predict appearance changes across seasons, *e.g.* summer to winter, Neubert *et al.* [52] use repeated recordings of the same scenes across time (captured by a moving train). Local patches (superpixels) are represented using visual vocabularies trained for each season where the visual words across different seasons are associated using the spatial layout of the images, *i.e.* the images across different seasons are assumed to be aligned with each other. In contrast, we focus on the place recognition across significant

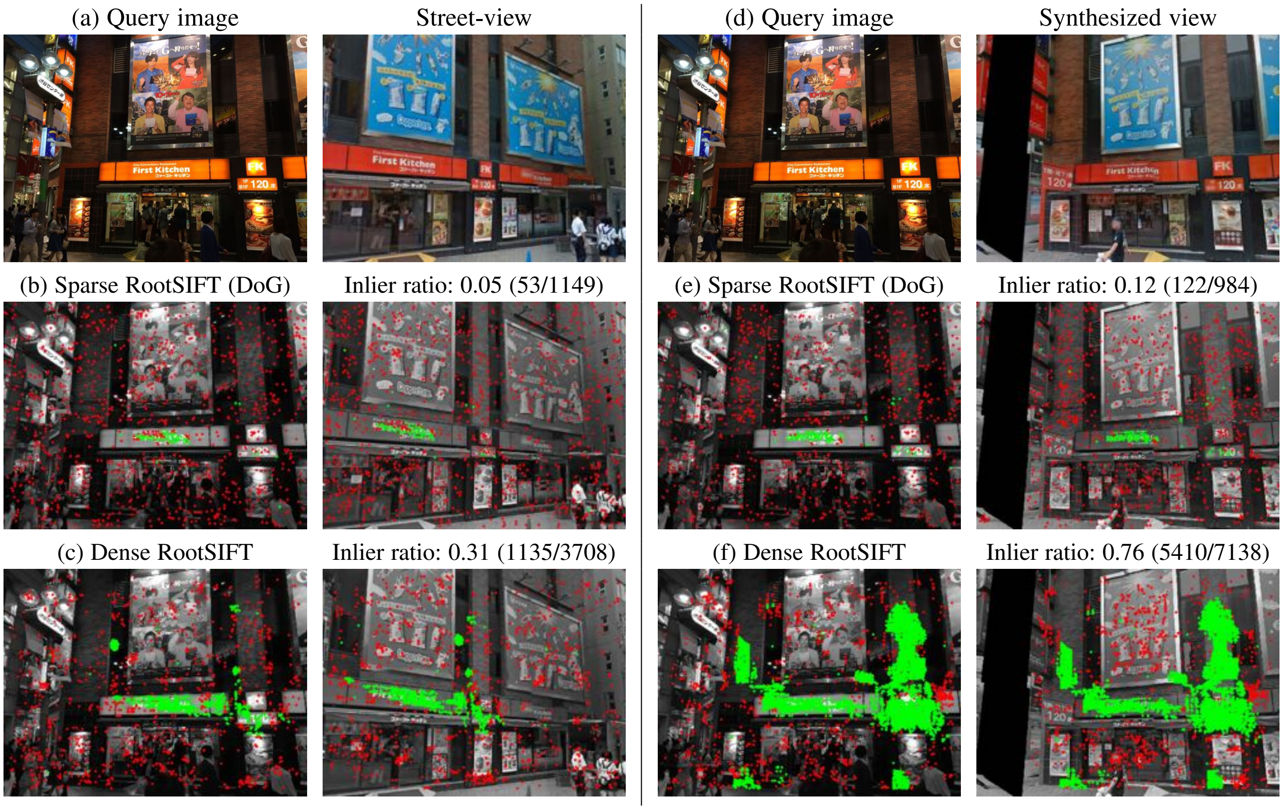

(a) Query image
Street-view
(d) Query image
Synthesized view
(b) Sparse RootSIFT (DoG)
Inlier ratio: 0.05 (53/1149)
(e) Sparse RootSIFT (DoG)
Inlier ratio: 0.12 (122/984)
(c) Dense RootSIFT
Inlier ratio: 0.31 (1135/3708)
(f) Dense RootSIFT
Inlier ratio: 0.76 (5410/7138)

Fig. 2. **Matching across illumination and structural changes in the scene. First row:** The same query image is matched to a street-view image depicting the same place from a different viewpoint (a) and to a synthesized virtual view depicting the query place from the same viewpoint (d). **Second row:** Matching sparsely sampled RootSIFT descriptors across a major change in illumination is difficult for the same (e) as well as for the different (b) viewpoints. **Third row:** Densely sampled descriptors can be matched across a large change in illumination (c) and the matching is much easier when the viewpoint is similar (f). In all cases the tentative matches are shown in red and geometrically verified matches are shown in green. Note how the proposed method (f), based on densely sampled descriptors coupled with virtual view synthesis, obtains significantly higher inlier ratio (0.76) on this challenging image pair with major illumination and structural changes in the scene.

changes of camera viewpoint as well as more severe appearance changes that comprise both different seasons and day/night illumination. Convolutional neural network (CNN) descriptors have been recently used for the robot re-localization task in [53] who extract object patches across the image and represent them by concatenating features from convolutional layers (AlexNet conv3) followed by random Gaussian projection demonstrating increased robustness to changes in illumination.

This paper is an extended version of [54] with complete description of our place recognition pipeline including detailed descriptions of several important components (new section 5) and several new experimental results including results on an new extended version of the Tokyo 24/7 image database (section 6).

3 Matching Local Descriptors Across Large Changes in Appearance

In this section we investigate the challenges of using local invariant features for image matching across major changes in scene appearance due to day/night illumination and structural changes in the scene. We first illustrate that local invariant features based on the difference of Gaussian (DoG) feature detector are not reliably repeatable in such conditions. Then we show that densely sampled descriptors result in better matches, but suffer from limited invariance to geometric transformations (scale and viewpoint). Finally, we demonstrate that matching can be significantly improved when we match to a virtual view synthesized from approximately the same viewpoint. In this section we illustrate the above points on a matching example shown in figure 2. We verify these findings quantitatively on the place recognition task in section 6.

In all examples in figure 2 we build tentative matches by finding mutually nearest descriptors. The tentative matches are shown in red. We then geometrically verify the matches by repeatedly finding several homographies using RANSAC. The geometrically consistent matches (inliers) are shown in green. We deem all geometrically verified matches as correct (though few incorrect matches may remain). The quality of matching is measured by the inlier ratio, *i.e.* the proportion of geometrically consistent matches in all tentative matches. The inlier ratio is between 0 and 1 with a perfect score of 1 when all tentative matches are geometrically consistent.

First, we match the upright RootSIFT descriptors [27] sampled at DoG keypoints [18] between a query image and a street-view image depicting the query place (figure 2(a)) from a different viewpoint. The matches are shown in figure 2(b) and result in inlier ratio of only 0.05, clearly demonstrating the difficulty of matching DoG keypoints across large changes in appearance.

Second, we repeat the same procedure for the synthesized view (figure 2(d)), which captures the query place from approximately the same viewpoint as the query image. The result is shown in figure 2(e). The resulting inlier ratio of only 0.12 indicates that matching the DoG keypoints across large changes in appearance is difficult despite the fact that the two views have the same

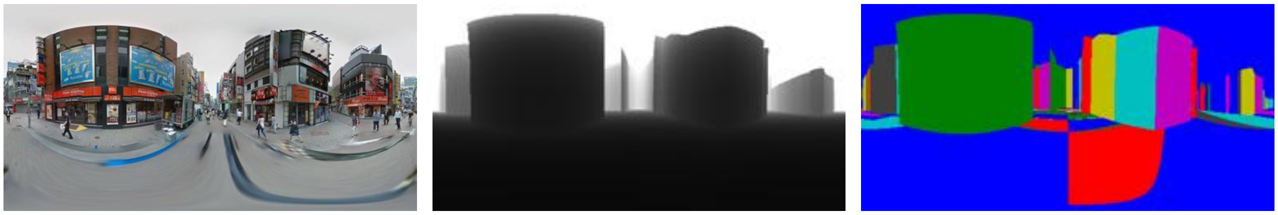

(a) Street-view panorama (b) Associated depth-map (c) Individual scene planes

Fig. 3. **Input data for view synthesis**. (a) The street-view panorama. (b) The associated piece-wise planar depth-map. Brightness indicates distance. (c) The individual scene planes are shown in different colors.

viewpoint.

Third, we extract RootSIFT descriptors with a width of 40 pixels (in a 640×480 image) on a densely sampled regular grid with a stride of 2 pixels. The descriptor matching was performed in the same manner as for the descriptors extracted at the sparsely detected keypoints. Matching the densely sampled descriptors across different viewpoints and illuminations already shows an improvement compared to sparse keypoints, with the inlier ratio increasing from 0.05 to 0.31 (figure 2(c)). The fact that the descriptor (RootSIFT) is identical for both sampling methods suggests that the main problem is non-repeatability of the Difference of Gaussian local invariant features underpinning the sparsely sampled method, rather than the descriptor itself.

Finally, we apply the densely sampled descriptors to the image pair with different illuminations but similar viewpoints (figure 2(d)). The matches are shown in figure 2(f). The inlier ratio further increases to 0.76 clearly demonstrating the benefits of virtual view synthesis for dense descriptor matching.

4 View synthesis from street-level imagery

In this section we describe our view synthesis method that expands the database of the geo-tagged images with additional viewpoints sampled on a regular grid. To synthesize additional views, we use the existing panoramic imagery together with an approximate piece-wise planar depth map associated with each panorama, as illustrated in figure 3. The piece-wise planar depth map provides only a very coarse 3D structure of the scene, which often leads to visible artifacts in the synthesized imagery. However, we demonstrate in section 6 that this quality is sufficient to significantly improve place recognition performance. In addition, this data is essentially available world-wide [55], thus opening up the possibility of planet-scale view synthesis and place recognition [24]. The view synthesis proceeds in two steps. We synthesize the candidate virtual camera locations, which is followed by synthesizing individual views. The two steps are discussed next.

We generate candidate camera positions on a regular $5m \times 5m$ grid on the map that covers the original street-view camera positions. We only generate camera positions that are within $20m$ distance from the original street-view trajectory, where the trajectory is obtained by connecting the neighboring street-view camera positions. We found that going farther than $20m$ often produces significant artifacts in the synthesized views. We also use the available depth maps to discard camera positions that would lie inside buildings. The camera positions of the synthesized views are illustrated on the map in figure 4.

To synthesize the virtual views at the particular virtual camera position we use the panorama and depth map downloaded from Google maps [55]. Each panorama captures $360°$ by $180°$ horizontal and vertical viewing angle, respectively, and has the size $13,312 \times 6,656$ pixels, as illustrated in figure 3(a). The depth map is encoded as a set of 3D plane parameters (normal and distance for each plane) and an 512×256 image of indices pointing, for each pixel, to one of the planes, as illustrated in figure 3(c). Using this index we can look-up the corresponding plane for each pixel, which allows us to generate the actual depth map for the panorama, as illustrated in figure 3(b). All views at a particular virtual camera position are synthesized from the panorama and depth map of the closest street-view image. Virtual views are synthesized by standard ray tracing with bilinear interpolation. In detail, for every pixel in the synthesized virtual view, we cast a ray from the center of the virtual camera, intersect it with the planar 3D structure obtained from the depth map of the closest street-view panorama, project the intersection to the street-view panorama, and interpolate the output pixel value from the neighboring pixels. For each virtual camera location we generate 12 perspective images of $1,280 \times 960$ pixels (corresponding to 60 degrees of horizontal field of view) with a pitch direction $12°$ and the following 12 yaw directions $[0°, 30°, ..., 360°]$. This perspective view sampling is similar to e.g. [4], [15]. Examples of the synthesized virtual views are shown in figures 1, 14 and 15. While the synthesized views have missing information and artifacts (e.g. incorrectly rendered people or objects), we found that this simple rendering is already sufficient to improve place recognition performance. Higher quality synthesis could be potentially obtained by combining information from multiple panoramas. Rendering one virtual view takes about a second, but we expect 1-2 orders of magnitude speed-up using a graphics processing unit (GPU). We generate the same set of perspective views for original street-view images and combine the real and virtual views into a single place recognition database. Note that virtual views are only needed for extracting the compact dense VLAD descriptors as described in section 3 and can be discarded afterwards.

5 Place recognition with view synthesis

In this section, we describe our complete place recognition method, illustrated in figure 6. Our pipeline has an off-line and an on-line stage. In the off-line stage, we generate a set of virtual camera poses, the corresponding perspective views and compute their densely extracted descriptors, thus significantly expanding the image database as described in section 4. In the on-line stage, we extract the dense descriptor from the query image (section 3), match it to the expanded database and retrieve the GPS positions of the top matches. Next we describe several important components of our system that allows us to perform these operations at large scale.

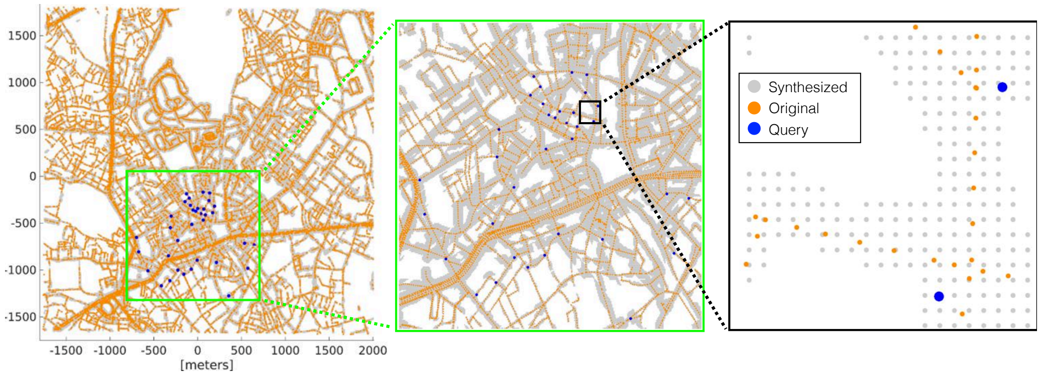
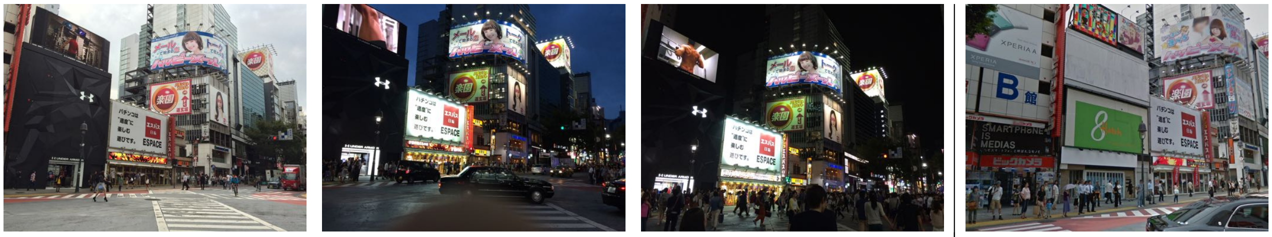


Fig. 4. **Combining street-view imagery with synthetic views.** The figure shows camera positions for part of the 24/7-Tokyo dataset. The positions of the original street-view images are shown in orange, the positions of synthesized views (5×5m grid) are shown in grey, and the positions of query images are shown in blue. The "large 24/7-Tokyo" database of geo-tagged images (left) includes $374,676$ views generated from $31,223$ street-view panoramas and $2,499,816$ synthesized views generated at $208,318$ virtual camera locations. The area bounded by green rectangle (middle) indicates the subset of the 24/7 large database that includes $75,984$ views generated from $6,332$ street-view panoramas. The inset (right) shows a close-up of one road intersection.



(a) Query 1. (b) Query 2. (c) Query 3. (d) Database image

Fig. 5. **Example query images from the newly collected 24/7 Tokyo dataset.** Each place in the query set is captured at different times of day: (a) daytime, (b) sunset, and (c) night. For comparison, the database street-view image at a close-by position is shown in (d). Note the major changes in appearance (illumination changes in the scene) between the database image (d) and the query images (a,b,c).

Dense VLAD in the expanded image database. As observed in section 3, matching images across large changes in appearance becomes possible thanks to densely sampled local features extracted from images with similar viewpoints. We implement this idea at scale by aggregating the densely extracted features into a single compact descriptor for an image. In detail, we describe each image by VLAD descriptor [31] that aggregates densely sampled RootSIFT descriptors [27], [32]. The dense VLAD descriptor is extracted from all the images in the expanded database.

Compression using product quantization. The image database expanded by a $5m \times 5m$ grid of synthetic views has 6 to 8 times larger memory footprint. To reduce these increased memory requirements we compress the extracted descriptors using product quantization (PQ) [21], [56], [57]. As will be shown in experiments in section 6 this achieves about an order of magnitude smaller representation with a negligible loss in place recognition accuracy.

Shortlist diversity. At query time, the dense VLAD descriptor is extracted from the query image and matched to the dense VLAD descriptors extracted offline from the entire database. The outcome is a short-list of top N matches for the query. One drawback of adding synthetic views is that the returned shortlist might be occupied by similar views from close by virtual viewpoints. Diversifying the returned shortlists [2] is, therefore, particularly important for our method. We diversify the returned shortlists using the fact that (synthesized) perspective views are generated from a much smaller set of street-view panoramic images. In detail, we first group both the synthetic and non-synthetic perspective views that are generated from the same street-view panoramic image. Then, we take only the best matching view in each group and add it into the shortlist.

6 Experiments

In this section we describe the newly collected 24/7 Tokyo dataset, give the place recognition performance measures and outline the quantitative and qualitative results of our method compared to several baselines.

6.1 Experimental setup

24/7 Tokyo dataset. We have collected a new test set of $1,125$ query images captured by Apple-iPhone5s and Sony-Xperia smartphones. We captured images at 125 distinct locations. At each location we captured images at 3 different viewing directions and at 3 different times of day, as illustrated in figure 5. The ground truth GPS coordinates at each location were recorded by manually localizing the position of the observer on the map at the finest zoom level. We estimate that the error of the ground truth location is below $5m$. The dataset is available at [58]. In the following evaluation, we use a subset of 315 query images within the area of about $1,600m \times 1,600m$ (the green rectangle in figure 4)).

We have constructed two geo-tagged image databases from Google street-view panoramas downloaded within the Tokyo metropolitan area. The larger database covers an area of about

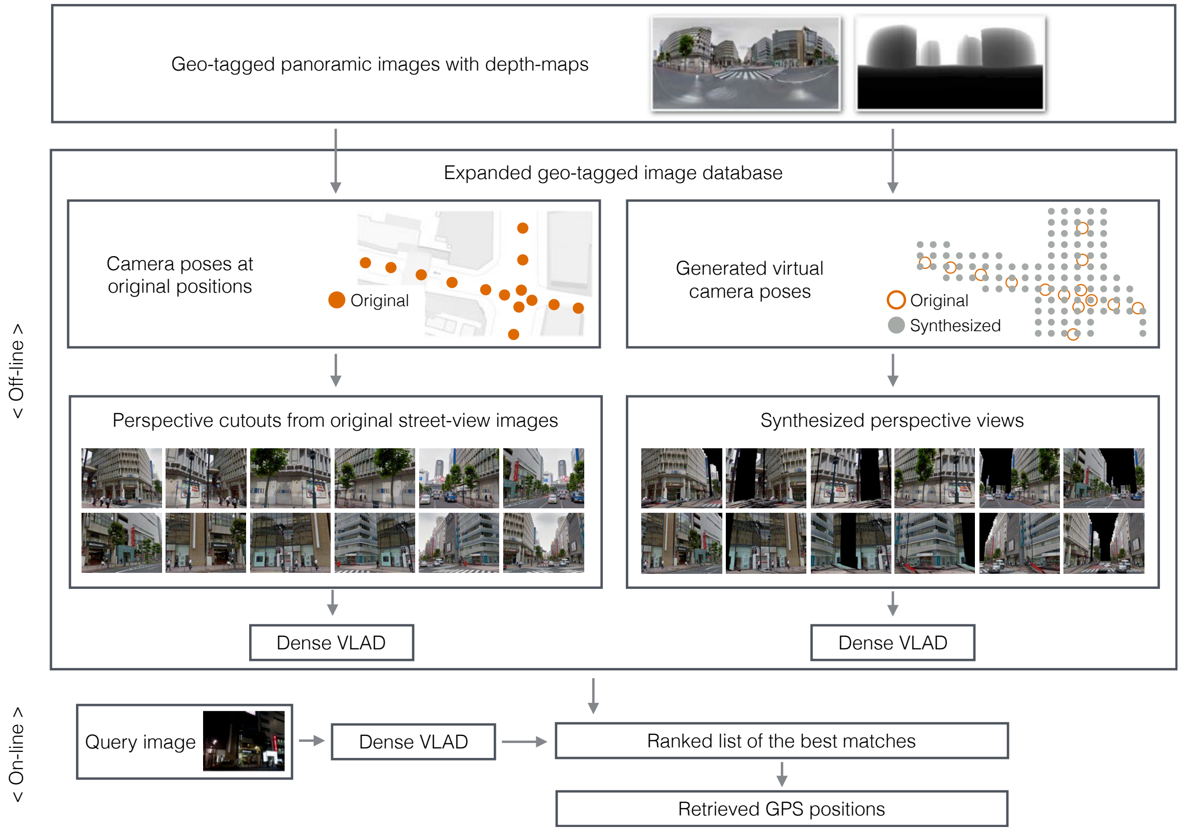

Geo-tagged panoramic images with depth-maps
Expanded geo-tagged image database
< Off-line >
Camera poses at original positions
Original
Generated virtual camera poses
Original
Synthesized
Perspective cutouts from original street-view images
Synthesized perspective views
Dense VLAD
Dense VLAD
< On-line >
Query image
Dense VLAD
Ranked list of the best matches
Retrieved GPS positions

Fig. 6. **Overview of our approach for place recognition with view synthesis.** Our approach has an off-line and an on-line stage. The off-line process generates expanded geo-tagged image database that includes Dense VLAD descriptors computed from perspective images of the original street-view images as well as the synthesized views. The position of the query image is estimated in the on-line stage by matching its Dense VLAD descriptor to descriptors in the expanded image database and retrieving the GPS positions of the top matches.

TABLE 1
Number of (perspective) images in the 24/7 Tokyo geotagged database. Number of original street-view panoramas is indicated in brackets.

	Large		Small	
Street-view	$374,676$	$(31,223)$	$75,984$	$(6,332)$
Synthesized views	$2,499,816$	$(208,318)$	$597,744$	$(49,812)$

$3,700m \times 3,700m$ (the entire area in figure 4, left). The smaller database corresponds to a sub-area of about $1,600m \times 1,600m$ (the green rectangle in figure 4). The number of images in each database is given in table 1. Unless stated otherwise, the smaller database is used for most of the comparative experiments in the paper. The large database is used to test the issues related to the scalability of our method.

Evaluation metric. The query image is deemed correctly recognized at N if at least one of the top N retrieved database images is within $d = 25$ meters from the ground truth position of the query. This is a common place recognition metric used in *e.g.* [4], [11], [15]. The percentage of correctly recognized queries (Recall) is then plotted for different values of N.

Implementation details. To compute the Dense VLAD descriptor, we re-size each image to have the maximum dimension of 640 pixels. This is beneficial for computational efficiency and limits the smallest scale of the extracted descriptors. We extract SIFT [18] descriptors at 4 scales corresponding to region widths of 16, 24, 32 and 40 pixels. The descriptors are extracted on a densely sampled regular grid with a stride of 2 pixels. When using synthesized images, we remove descriptors that overlap with image regions that have no image data (shown in black in the synthesized imagery). We use the SIFT implementation available in Vlfeat [59] followed by the RootSIFT normalization [27], *i.e.* L1 normalization followed by element-wise square root. The visual vocabulary of 128 visual words (centroids) is built from 25M descriptors randomly sampled from the database images using k-means clustering[1]. We have kept the original dimension of the SIFT descriptor, unlike [31]. Each image is then described by an aggregated intra-normalized [25] VLAD descriptor followed by a PCA compression to 4,096 dimensions, whitening and L2 normalization [60]. Similarity between the test query and each database image is measured using the normalized dot product between their descriptors.

6.2 Comparison with baseline methods

In this section we compare performance of our approach to a number of baseline methods. We then evaluate benefits of using higher image resolution.

Baseline methods. We compare results to the following baselines.

- **Sparse VLAD.** Here upright RootSIFT descriptors extracted at the Difference of Gaussian (DoG) local invariant

[1] Please note that the results presented in this journal version are slightly different from those in the conference version as we have re-implemented the view synthesis and the extraction of the Dense VLAD descriptors, as well as built a new set of visual vocabularies.

TABLE 2
Comparison with the baseline methods on the 24/7-Tokyo dataset using 4,096 dimensional image descriptors.

Method	Recall			
	top 1	top 10	top 20	top 50
Dense VLAD SYNTH (our)	66.03	75.87	80.32	85.08
Dense VLAD	60.95	72.06	74.60	80.00
Sparse VLAD	25.71	41.90	47.94	57.46
Dense FV	54.29	71.11	74.60	77.78
Sparse FV	41.59	59.05	65.40	73.65

features [18], [59] are aggregated into a single VLAD descriptor. Apart from changing the image sampling (from dense to sparse DoG) the descriptor is extracted in the same manner as our densely sampled VLAD.

- **Sparse FV.** Second, we compare with the standard sparse Fisher vector [31] (Sparse FV), which was shown to perform well for place recognition [15]. The Fisher vector is constructed using the same upright RootSIFT descriptors as the Sparse VLAD baseline. Following [31], extracted RootSIFT descriptors are reduced to 64 dimensions by PCA. A 256-component Gaussian mixture model is then trained from 25M descriptors randomly sampled from the database images. As in [31], the resulting 256×64 dimensional Fisher vector is reduced to $4,096$ dimensions using PCA, followed by whitening and L2 normalization [60].
- **Dense FV.** Third, we evaluate the Fisher vector based on densely detected features. We use the same descriptors and normalization as used in our Dense VLAD to evaluate impact of the different aggregation schemes. Using the same 25M of 128 dimensional RootSIFT descriptors detected on a densely sampled regular grid, we train a 128-component Gaussian mixture model. Each image is then described by an intra-normalized FV descriptor followed by a PCA compression to $4,096$ dimensions, whitening and L2 normalization.
- **Sparse BoVW** Finally, we also compare results to the bag-of-visual-words baseline. We construct the bag-of-visual-words descriptor (Sparse BoVW) using the same upright RootSIFT descriptors as used in the Sparse VLAD baseline. A vocabulary of $200,000$ visual words is built by approximate k-means clustering [19], [61]. The resulting bag-of-visual-word vectors are re-weighted using the adaptive assignment described in [15].

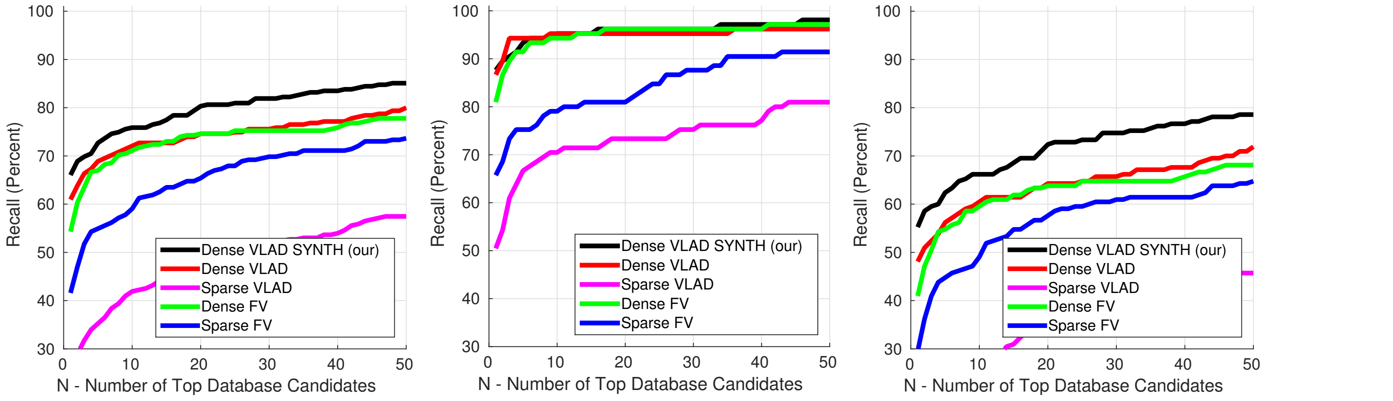
Note that we focus on comparing with other compact image representations (VLAD / Fisher vectors) and do not compare here to methods that store local-invariant features for each image, such as [1], which requires $40\times$ more memory per image, and $6\times$ more memory in total (accounting for the $6.7\times$ increase in the number of images due to the novel views).

Benefits of the dense descriptor and synthesized views. First, in figure 7 and table 2 we evaluate the benefits of having dense descriptors (Dense VLAD, Dense FV). We compare performance to the standard VLAD and Fisher vector descriptors sampled at local invariant features (Sparse VLAD, Sparse FV). We show results for all queries (figure 7(a)), but to clearly illustrate the differences we also separate the query images into daytime (figure 7(b)), and sunset/night queries (figure 7(c)). Results clearly demonstrate that dense descriptors (Dense VLAD, Dense FV) improve over the sparse baselines (Sparse VLAD, Sparse FV). We next evaluate the benefits of having additional synthesized views (Dense VLAD SYNTH). While having the dense descriptor (Dense VLAD) already improves performance, it is the combination of the dense descriptor with synthetic virtual views (Dense VLAD SYNTH) which brings improvements especially for queries with difficult illumination (figure 7(c)), clearly illustrating the importance of both components of our approach.

Comparison to sparse baselines. In figure 8, we separately show a comparison of our method (Dense VLAD SYNTH) to several baselines that use only sparsely sampled local invariant features. Overall, our method significantly improves over all sparse baselines. Further analysis reveals that for VLAD computed from (sparse) DoG keypoints, adding synthetic virtual views (Sparse VLAD SYNTH) helps (compared to Sparse VLAD). In contrast, adding synthetic virtual views to Fisher vector matching (Sparse FV SYNTH) does not improve over the standard FV without virtual views (Sparse FV).

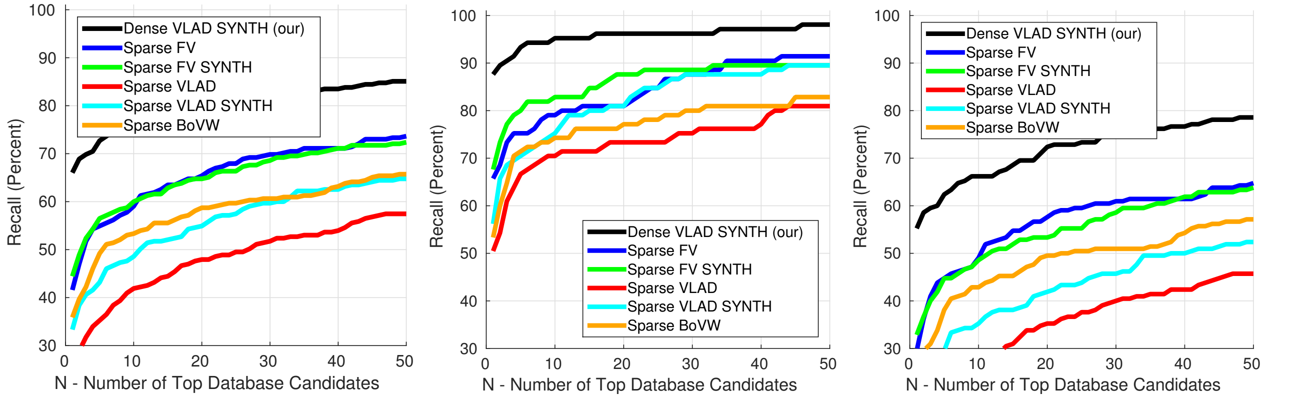
Benefits of view synthesis for CNN-based image descriptor (NetVLAD) [42]. We next evaluate the benefits of view synthesis for the recent NetVLAD descriptor [42]. This descriptor is based on a convolutional neural network trained in an end-to-end manner for the place recognition task using a weakly supervised triplet loss on Google Street-view time machine imagery. This is a strong very recent baseline achieving state-of-the-art results on place recognition benchmarks. Our initial investigation revealed that applying the NetVLAD descriptor out of the box on synthetic data can have a negative impact on performance (decreasing the recall@20 from 98.1% to 97.1% for day time queries). After further analysis we found that this can be attributed to the view synthesis artifacts (see the large areas of black pixels in the second column of figure 14). In the case of DenseVLAD, SIFT descriptors affected by synthesis artifacts can be easily removed as SIFT has only a limited extent in the image. However, this is not the case for NetVLAD where the receptive field of the aggregated conv5 features is large and the artifacts are often affecting all conv5 features across the image. To address this issue, we have re-trained the NetVLAD descriptor with images having missing pixels with the hope that the network will learn to ignore those artifacts in a similar way it learns to ignore transient objects such as cars and people that are not informative for identifying a specific place [42]. In detail, we have extracted masks of missing pixels from our synthetic views and overlaid them in a random manner over the TokyoTM training dataset from [42]. The descriptor was then re-trained in the same manner as in [42]. Results are shown in figure 9 and clearly demonstrate the benefits of synthetic views for this re-trained NetVLAD descriptor. Note, however, that training with missing pixels lowers slightly the absolute performance of the method compared to the original NetVLAD descriptor. Removing the artifacts altogether, for example by using multiple images [49], [50], [62] combined with hole-filling [63], is likely to further improve the results.

Benefits of higher image resolution. In figure 10, we investigate how the place recognition performance of the Dense VLAD descriptor changes when high resolution images are available for both the query and database. In detail, we use high-resolution images that have maximum dimension of 1280 pixels (compared to 640 pixels used in the rest of the paper). Each high-res image is described by Dense VLAD descriptor (Dense VLAD 1280×960, Dense VLAD SYNTH 1280×960) in the same parameter setup as



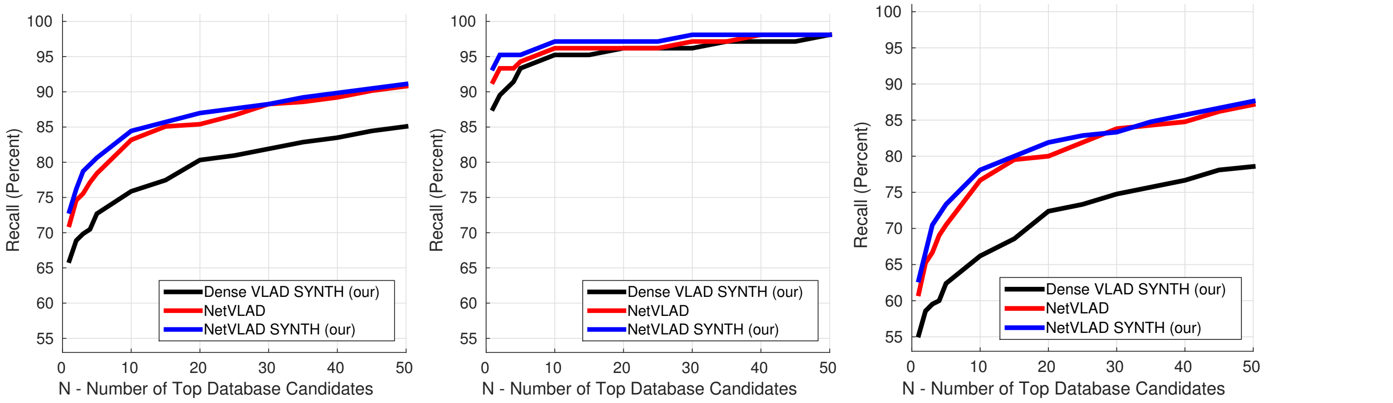
(a) All queries (b) Day time queries (c) Sunset and night queries

Fig. 7. **Evaluation on the 24/7-Tokyo dataset**. The fraction of correctly recognized queries (Recall, y-axis) vs. the number of top N retrieved database images (x-axis) for the proposed method (Dense VLAD SYNTH) compared to the baseline methods (Dense VLAD, Dense FV, Sparse FV). The performance is evaluated for all test query images (a), as well as separately for daytime queries (b), and sunset/night queries (c). The benefits of the proposed method (Dense VLAD SYNTH) is most prominent for difficult illuminations (c).



(a) All queries (b) Day time queries (c) Sunset and night queries

Fig. 8. **Comparison with the baseline methods on the 24/7-Tokyo dataset.** Each plot shows the fraction of correctly recognized queries (Recall, y-axis) vs. the number of top N retrieved database images (x-axis).



(a) All queries (b) Day time queries (c) Sunset and night queries

Fig. 9. **Benefits of synthetic views for the the NetVLAD descriptor [42] on the 24/7-Tokyo dataset.** Using synthetic views (NetVLAD SYNTH) benefits also the state-of-the-art CNN-based descriptor (NetVLAD). For both NetVLAD and NetVLAD SYNTH the descriptor has been trained on images with missing pixels to learn to ignore artifacts in view synthesis. For comparison, place recognition performance for the proposed method using densely extracted RootSIFT descriptors (Dense VLAD SYNTH) is shown in black.

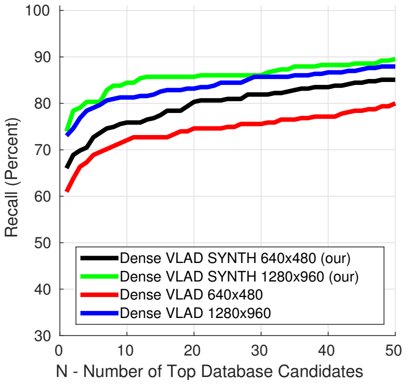


Fig. 10. **Benefits of higher image resolution.** The plot shows the fraction of correctly recognized queries (Recall, y-axis) vs. the number of top N retrieved database images (x-axis).

the low-res images. Notice that using higher resolution images requires additional computational time for RootSIFT extraction and aggregation but no additional space/time for the image descriptor matching as the dimensionality of the descriptor remains the same. Interestingly, we observe a significant gain in performance for both the Dense VLAD and Dense VLAD SYNTH methods. This suggests that high resolution images should be used whenever available.

6.3 Scaling-up to city-scale datasets

In this section we investigate several issues related to scaling-up the proposed method to large image datasets. First, we study how the place recognition performance decreases with dataset size. Second, we investigate techniques for compressing image descriptors. Third, we study the effect of descriptor dimensionality on the place recognition performance and investigate the effect of reducing the number of synthesized views. Finally, we present the qualitative results and outline the limitations of our method. Details are given next.

Scalability. We evaluate the scalability of our method on the large Tokyo 24/7 database (figure 4, table 1) that has 4.9 more images than the small database used in the experiments so far. We use the same test query set as for the small database in order to investigate the change in performance with the increased database size. We choose the standard (sparse) Fisher vector descriptor as the baseline method as it was found to work well for place recognition in other work [15] and performed well on the small database. The results are shown in figure 11 and exhibit a similar pattern as for the small database (figure 7): Dense VLAD improves performance compared to the Sparse FV baseline and synthesizing virtual views (Dense VLAD SYNTH) brings an additional improvement. Overall, indexing the larger database with our method results in a small drop in performance (recall@1 going from 62.9% to 62.2%).

Descriptor compression. For the 24/7 Tokyo large dataset, our method synthesizes 2.5M virtual views compared to 375K perspective street-view images in the same area. Hence, our method needs to index about 6.7 times more images compared to baselines without virtual view synthesis. Scaling-up towards place recognition in the entire city can be achieved by compressing the extracted descriptors using Product Quantization (PQ) [21]. In figure 11, we evaluate how the PQ compression changes the place recognition performance for the dense descriptors and additional synthesized views. Product Quantization with 4-D blocks encoded with $4,096$ centroids (12 bits) achieves $10.7\times$ compression with a negligible loss in accuracy (Dense VLAD PQ vs. Dense VLAD SYNTH PQ). Note that our method with synthesized views and compressed descriptors (Dense VLAD SYNTH PQ) requires a smaller memory than the original Dense VLAD but achieves a better place recognition accuracy.

Analysis of descriptor dimensionality. In figure 12 we investigate how the place recognition performance changes with reducing the dimensionality of the Dense VLAD descriptor from $4,096$ to $2,048$, $1,024$ and 512 dimensions. We observe a drop in performance especially for the lowest dimension. This suggests, that having a sufficiently rich representation is important for matching across large changes in appearance.

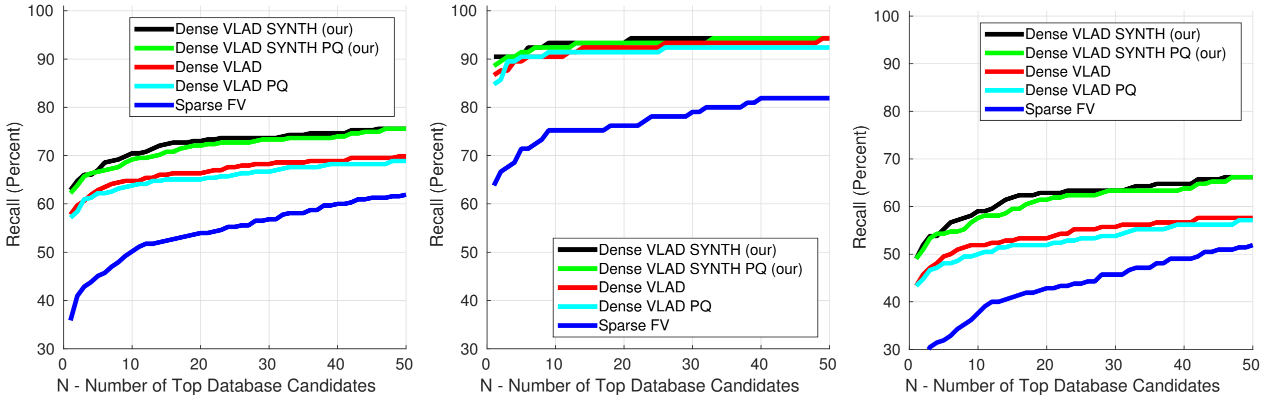
How many virtual views? In figure 13 we evaluate the required sampling of virtual views. First, we subsample the virtual views spatially from 5×5 meter grid (used in our method so far) to 10×10 meter grid. The spatial subsampling to 10×10 can reduce the number of virtual views by 75% with only a relatively small drop in place recognition performance. Then we subsample the number of yaw directions to only 6 per camera position, one every 60° (Dense VLAD SYNTH 60deg) compared to 12 yaw directions, one every 30° used in our method. In this experiment we keep the spatial sampling to 5×5 meters. Although the angular subsampling reduces the number synthetic views by only 50% it results in a fairly significant drop in performance, especially at the top 1 position. Note that in some situations a prior information on typical locations of query images could be available. For example pictures taken by pedestrians are likely to be taken on side-walks and imagery from vehicle-mounted cameras is likely to be captured on the streets. In these situations, we can optimize where to synthesize virtual views or, alternatively, we can pre-compute all virtual views in a dense manner and consider only a subset of descriptors relevant to the query on the fly.

Qualitative results. Figures 14 and 15 show examples of place recognition results. Notice that query images (left column) include large changes in both viewpoint and illumination compared to the available street-view for the same places (right column). The synthesized views (2nd column) at new positions significantly reduce the variation in viewpoint and thus enable matching across large illumination changes, as discussed in section 3.

Limitations. Figure 16 shows examples of queries which remain very difficult to localize. The typical failure modes are (i) very dark night time images with limited dynamic range, (ii) places with vegetation, which is hard to uniquely describe using the current representation, and (iii) places where view synthesis fails often due to complex underlying 3D structure not captured well by the approximate depth maps available with street-view imagery.

7 Conclusion

We have described a place recognition approach combining synthesis of new virtual views with a densely sampled but compact image descriptor. The proposed method enables true 24/7 place recognition across major changes in scene illumination throughout the day and night. We have experimentally shown its benefits on a newly collected place recognition dataset – *24/7 Tokyo* – capturing the same locations in vastly different illuminations. Our work is another example in the recent trend showing benefits of



(a) All queries (b) Day time queries (c) Sunset and night queries

Fig. 11. **Scaling-up to the large 24/7-Tokyo database**. The place recognition performance for the proposed method (Dense VLAD SYNTH, Dense VLAD SYNTH PQ) compared to the baseline methods (Dense VLAD, Dense VLAD PQ, Sparse FV).

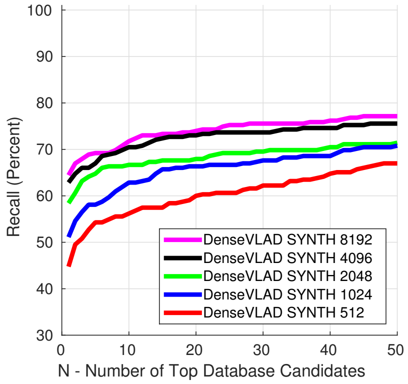


Fig. 12. **Evaluation of the dimensionality reduction on the large 24/7-Tokyo database**. The plot shows the fraction of correctly recognized queries (Recall, y-axis) vs. the number of top N retrieved database images (x-axis).

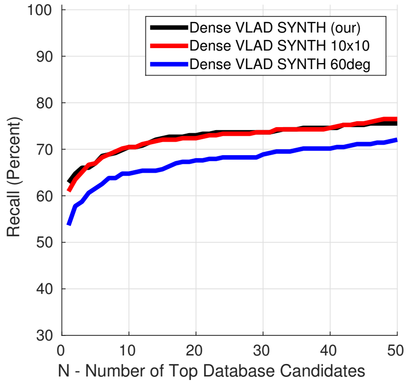


Fig. 13. **Evaluation of view sampling density on the large 24/7-Tokyo database**. The plot shows the fraction of correctly recognized queries (Recall, y-axis) vs. the number of top N retrieved database images (x-axis).

3D structure for visual recognition. As we build on the widely available Google street-view imagery our work opens-up the possibility of planet-scale 24/7 place recognition.

Acknowledgments

The authors were partially supported by JSPS KAKENHI Grant Number 15H05313, Project LADIO EU H2020 No. 731970, Google, CityLabs@Inria, the ERC project LEAP (no. 336845), ANR project Semapolis (ANR-13-CORD-0003), CIFAR Learning in Machines&Brains program, Project IMPACT CZ.02.1.01/0.0/0.0/15_003/0000468 and the European Regional Development Fund, and the Intelligence Advanced Research Projects Activity (IARPA) via Air Force Research Laboratory, contract FA8650-12-C-7212. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, AFRL, or the U.S. Government.

References

[1] R. Arandjelović and A. Zisserman, "DisLocation: Scalable descriptor distinctiveness for location recognition," in *Proc. Asian Conf. on Computer Vision*, 2014.

[2] S. Cao and N. Snavely, "Graph-Based Discriminative Learning for Location Recognition," in *CVPR*, 2013.

[3] ——, "Minimal Scene Descriptions from Structure from Motion Models," in *CVPR*, 2014.

[4] D. Chen, G. Baatz *et al.*, "City-scale landmark identification on mobile devices," in *CVPR*, 2011.

[5] M. Cummins and P. Newman, "Highly scalable appearance-only SLAM - FAB-MAP 2.0," in *Proceedings of Robotics: Science and Systems*, Seattle, USA, June 2009.

[6] J. Hays and A. Efros, "im2gps: estimating geographic information from a single image," in *CVPR*, 2008.

[7] J. Knopp, J. Sivic, and T. Pajdla, "Avoiding Confusing Features in Place Recognition," in *ECCV*, 2010.

[8] Y. Li, N. Snavely, D. Huttenlocher, and P. Fua, "Worldwide Pose Estimation Using 3D Point Clouds," in *ECCV*, 2012.

[9] N. Sünderhauf, M. Milford, P. Corke, and T. Sattler, "CVPR 2015 workshop on visual place recognition in changing environments," June 2015. [Online]. Available: https://roboticvision.atlassian.net/wiki/pages/viewpage.action?pageId=14188617

[10] T. Sattler, B. Leibe, and L. Kobbelt, "Improving Image-Based Localization by Active Correspondence Search," in *ECCV*, 2012.

[11] T. Sattler, T. Weyand, B. Leibe, and L. Kobbelt, "Image Retrieval for Image-Based Localization Revisited," in *BMVC*, 2012.

[12] T. Sattler, M. Havlena, F. Radenović, K. Schindler, and M. Pollefeys, "Hyperpoints and fine vocabularies for large-scale location recognition," in *ICCV*, 2015.

[13] T. Sattler, M. Havlena, K. Schindler, and M. Pollefeys, "Large-scale location recognition and the geometric burstiness problem," in *CVPR*, 2016.

[14] G. Schindler, M. Brown, and R. Szeliski, "City-Scale Location Recognition," in *CVPR*, 2007.

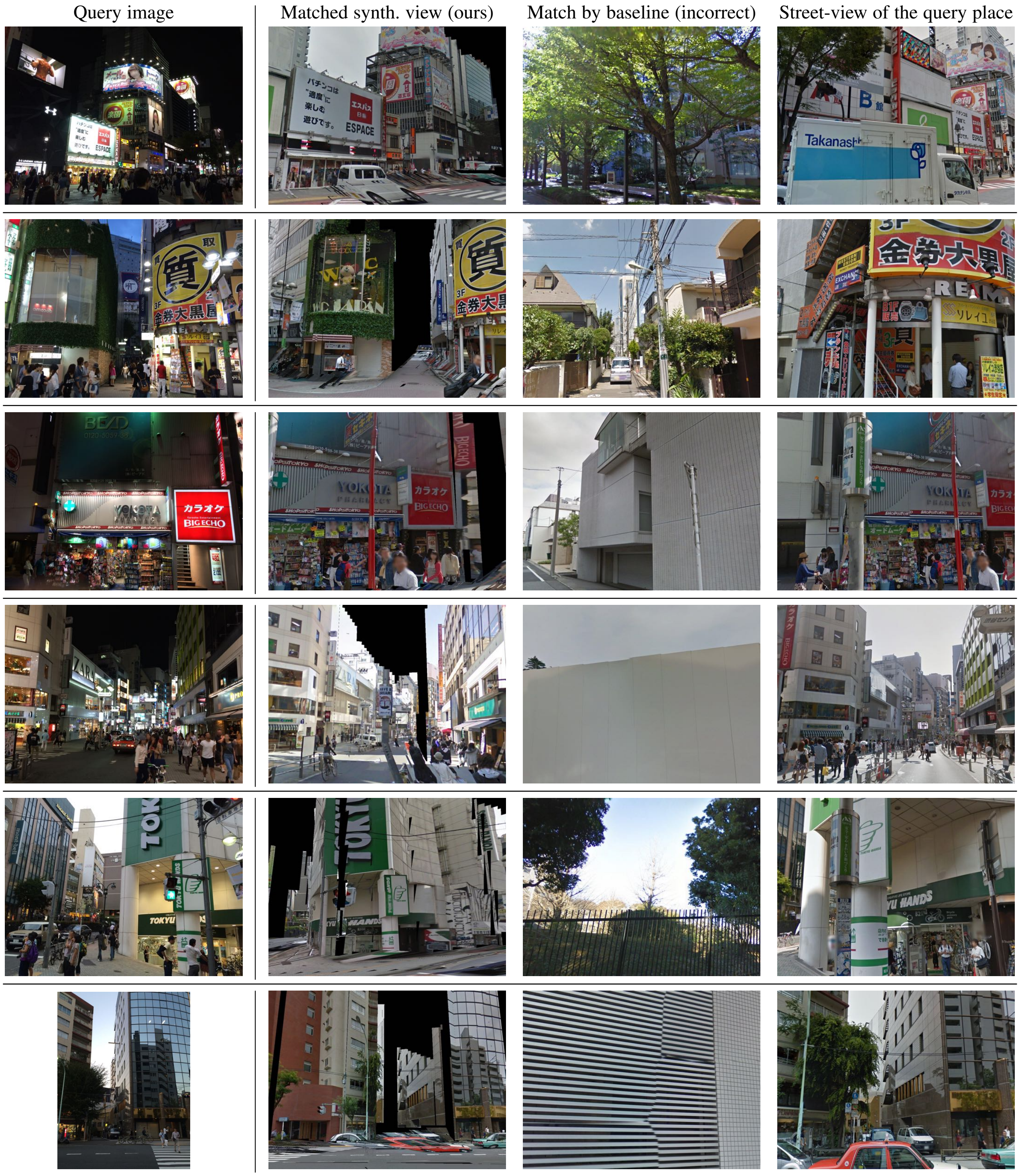

Query image
Matched synth. view (ours)
Match by baseline (incorrect)
Street-view of the query place

Fig. 14. **Example place recognition results for our method (Dense VLAD SYNTH) compared to baseline using only sparsely sampled feature points (Sparse FV).** (Left) Query image. (2nd column) The best matching synthesized view by our method (correct). (3rd column) The best matching street-view image by the baseline (Sparse Fisher vectors without synthesized views). (4th column) The original street-view image at the closest position to the query. Note that our method can match difficult queries with challenging illumination conditions.

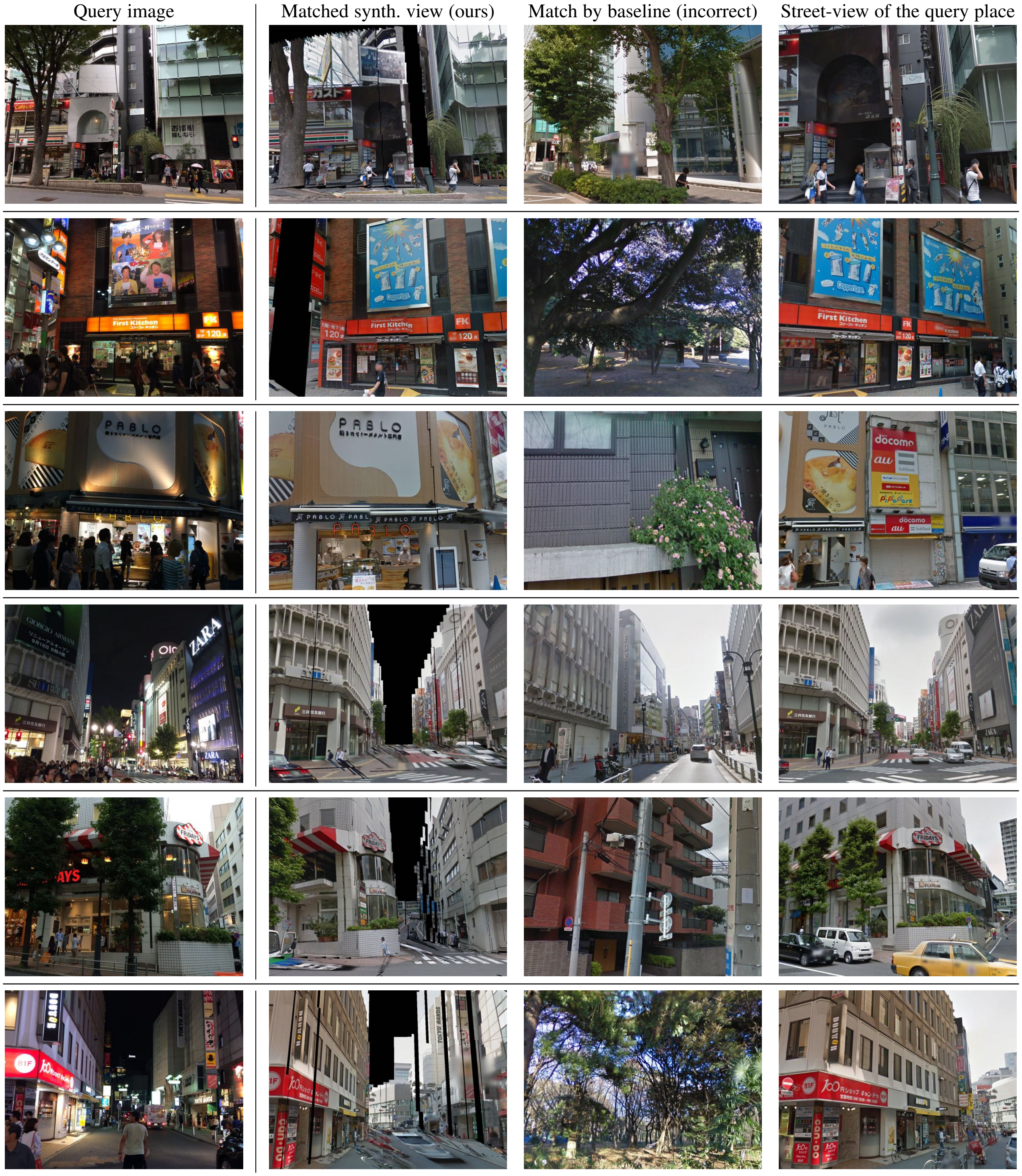

Query image
Matched synth. view (ours)
Match by baseline (incorrect)
Street-view of the query place

Fig. 15. **Example place recognition results with synthesized views (our method) compared to using only the original Google street-view images.** (Left) Query image. Note the difficult illumination. (2nd column) The best matching image (correct) by our method (Dense VLAD descriptor with the database expanded by synthesized views). (3rd column) The best matching image (incorrect) by Dense VLAD matching but using only the original street-view images. (4th column) The original street-view database image at the closest position to the query. Our method (2nd column) that uses virtual views with very similar viewpoints to the query can localize queries with difficult (night) illumination, thus enabling true 24/7 localization. This is not possible using the original street-view images (4th column), which depict the same places but from quite different viewpoints. **Please see additional results on the project webpage [58]**

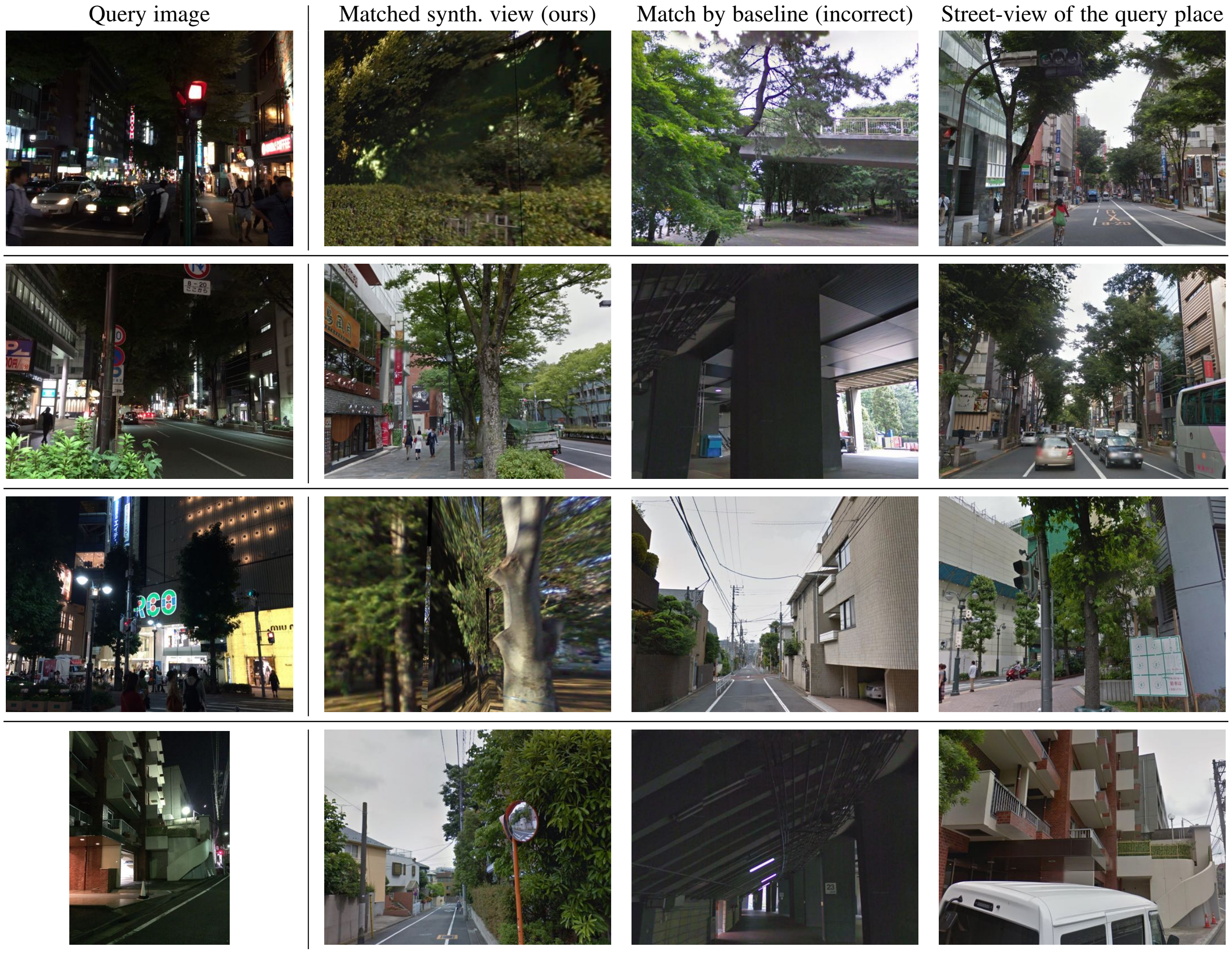

Query image
Matched synth. view (ours)
Match by baseline (incorrect)
Street-view of the query place

Fig. 16. **Examples of challenging query images that remain hard to localize.** The best match by our method (2nd column) and the street-view image by the baseline (Sparse Fisher vectors without synthesized views) (3rd column) both fail to correctly localize the query image (left). The original street-view image at the closest position to the query is shown in the 4th column.

[15] A. Torii, J. Sivic, T. Pajdla, and M. Okutomi, "Visual Place Recognition with Repetitive Structures," in *CVPR*, 2013.

[16] A. R. Zamir and M. Shah, "Accurate Image Localization Based on Google Maps Street View," in *ECCV*, 2010.

[17] F. Li and J. Kosecka, "Probabilistic location recognition using reduced feature set," in *Proc. Int. Conf. on Robotics and Automation*, 2006.

[18] D. Lowe, "Distinctive image features from scale-invariant keypoints," *IJCV*, vol. 60, no. 2, pp. 91–110, 2004.

[19] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in *CVPR*, 2007.

[20] J. Sivic and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos," in *ICCV*, 2003.

[21] H. Jégou, M. Douze, and C. Schmid, "Product quantization for nearest neighbor search," *PAMI*, vol. 33, no. 1, pp. 117–128, 2011.

[22] D. Hauagge and N. Snavely, "Image matching using local symmetry features," in *CVPR*, 2012.

[23] K. Matzen and N. Snavely, "Scene chronology," in *ECCV*, 2014.

[24] B. Klingner, D. Martin, and J. Roseborough, "Street view motion-from-structure-from-motion," in *ICCV*, 2013.

[25] R. Arandjelović and A. Zisserman, "All about VLAD," in *CVPR*, 2013.

[26] H. Jégou, M. Douze, C. Schmid, and P. Perez, "Aggregating local descriptors into a compact image representation," in *CVPR*, 2010.

[27] R. Arandjelović and A. Zisserman, "Three things everyone should know to improve object retrieval," in *CVPR*, 2012.

[28] T. Tuytelaars and K. Mikolajczyk, "Local invariant feature detectors: a survey," *Foundations and Trends in Computer Graphics and Vision*, vol. 3, no. 3, pp. 177–280, 2008.

[29] M. Aubry, B. C. Russell, and J. Sivic, "Painting-to-3d model alignment via discriminative visual elements," *ACM Transactions on Graphics (TOG)*, vol. 33, no. 2, p. 14, 2014.

[30] O. Chum, J. Philbin, J. Sivic, M. Isard, and A. Zisserman, "Total recall: Automatic query expansion with a generative feature model for object retrieval," in *ICCV*, 2007.

[31] H. Jégou, F. Perronnin, M. Douze, J. Sánchez, P. Pérez, and C. Schmid, "Aggregating local image descriptors into compact codes," *PAMI*, vol. 34, no. 9, pp. 1704–1716, 2012.

[32] A. Bosch, A. Zisserman, and X. Munoz, "Image classification using random forests and ferns," in *ICCV*, 2007.

[33] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection," in *CVPR*, 2005.

[34] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *CVPR*, 2006.

[35] A. Oliva and A. Torralba, "Modeling the shape of the scene: a holistic representation of the spatial envelope," *IJCV*, vol. 42, no. 3, pp. 145–175, 2001.

[36] W. Zhao, H. Jégou, and G. Gravier, "Oriented pooling for dense and non-dense rotation-invariant features," in *BMVC*, 2013.

[37] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural Computation*, vol. 1, no. 4, pp. 541–551, 1989.

[38] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[39] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "CNN features off-the-shelf: An astounding baseline for recognition," *CoRR*, vol. abs/1403.6382, 2014.

[40] H. Azizpour, A. Razavian, J. Sullivan, A. Maki, and S. Carlsson, "Factors

of transferability from a generic ConvNet representation," *CoRR*, vol. abs/1406.5774, 2014.

[41] A. S. Razavian, J. Sullivan, A. Maki, and S. Carlsson, "A baseline for visual instance retrieval with deep convolutional networks," in *Proc. ICLR*, 2015.

[42] R. Arandjelović, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "NetVLAD: CNN architecture for weakly supervised place recognition," in *CVPR*, 2016.

[43] A. Irschara, C. Zach, J. Frahm, and H. Bischof, "From structure-from-motion point clouds to fast location recognition," in *CVPR*, 2009.

[44] Q. Shan, C. Wu, B. Curless, Y. Furukawa, C. Hernandez, and S. M. Seitz, "Accurate geo-registration by ground-to-aerial image matching," in *3DV*, 2014.

[45] C. Wu, B. Clipp, X. Li, J.-M. Frahm, and M. Pollefeys, "3D model matching with viewpoint-invariant patches (VIP)," in *CVPR*, 2008.

[46] D. Sibbing, T. Sattler, B. Leibe, and L. Kobbelt, "SIFT-Realistic Rendering," in *3DV*, 2013.

[47] D. Hauagge, S. Wehrwein, P. Upchurch, K. Bala, and N. Snavely, "Reasoning about photo collections using models of outdoor illumination," in *BMVC*, 2014.

[48] N. Jacobs, S. Satkin, N. Roman, R. Speyer, and R. Pless, "Geolocating static cameras," in *ICCV*, 2007.

[49] P.-Y. Laffont, Z. Ren, X. Tao, C. Qian, and J. Hays, "Transient attributes for high-level understanding and editing of outdoor scenes," *ACM Trans. Graphics*, vol. 33, no. 4, 2014.

[50] K. Kim, A. Torii, and M. Okutomi, "Multi-view inverse rendering under arbitrary illumination and albedo," in *ECCV*, 2016.

[51] W. Maddern, A. Stewart, C. McManus, B. Upcroft, W. Churchill, and P. Newman, "Illumination invariant imaging: Applications in robust vision-based localisation, mapping and classification for autonomous vehicles," in *Proceedings of the Visual Place Recognition in Changing Environments Workshop, IEEE International Conference on Robotics and Automation (ICRA)*, Hong Kong, China, May 2014.

[52] P. Neubert, N. Sunderhauf, and P. Protzel, "Superpixel-based appearance change prediction for long-term navigation across seasons," *Robotics and Autonomous Systems*, vol. 69, pp. 15 – 27, 2015.

[53] N. Sunderhauf, S. Shirazi, A. Jacobson, F. Dayoub, E. Pepperell, B. Upcroft, and M. Milford, "Place recognition with convnet landmarks: Viewpoint-robust, condition-robust, training-free," in *Robotics: Science and Systems*, July 2015.

[54] A. Torii, R. Arandjelović, J. Sivic, M. Okutomi, and T. Pajdla, "24/7 place recognition by view synthesis," in *CVPR*, 2015.

[55] http://maps.google.com/help/maps/streetview/.

[56] M. Norouzi and D. Fleet, "Cartesian k-means," in *CVPR*, 2013.

[57] T. Ge, K. He, Q. Ke, and J. Sun, "Optimized product quantization," *PAMI*, vol. 36, no. 4, pp. 744–755, 2014.

[58] http://www.ok.ctrl.titech.ac.jp/∼torii/project/247/.

[59] A. Vedaldi and B. Fulkerson, "VLFeat: An open and portable library of computer vision algorithms," http://www.vlfeat.org/, 2008.

[60] H. Jégou and O. Chum, "Negative evidences and co-occurrences in image retrieval: the benefit of PCA and whitening," in *ECCV*, 2012.

[61] M. Muja and D. Lowe, "Fast approximate nearest neighbors with automatic algorithm configuration," in *VISAPP*, 2009.

[62] J. Flynn, I. Neulander, J. Philbin, and N. Snavely, "Deepstereo: Learning to predict new views from the world's imagery," in *CVPR*, 2016. [Online]. Available: http://www.cv-foundation.org/openaccess/content_cvpr_2016/html/Flynn_DeepStereo_Learning_to_CVPR_2016_paper.html

[63] D. Pathak, P. Krähenbühl, J. Donahue, T. Darrell, and A. Efros, "Context encoders: Feature learning by inpainting," in *CVPR*, 2016.

Akihiko Torii Akihiko Torii received a Master degree and PhD from Chiba University in 2003 and 2006. He then spent four years as a post-doctoral researcher in Czech Technical University in Prague. Since 2010, Since 2010, he has been with Tokyo Institute of Technology, where he is currently an assitant professor in the Department of Systems and Control Engineering, the School of Engineering.

Relja Arandjelović Relja Arandjeloviić received the BA and MEng degrees from the University of Cambridge in 2009, and PhD from the University of Oxford in 2013. He then spent one year as a postdoctoral researcher at the University of Oxford, and two years as a postdoctoral researcher at INRIA / Ecole Normale Superieure, Paris. Since 2016, he is a senior research scientist at DeepMind. His research interests include large-scale visual search and place recognition.

Josef Sivic Josef Sivic received a degree from the Czech Technical University, Prague, in 2002 and PhD from the University of Oxford in 2006. His thesis dealing with efficient visual search of images and videos was awarded the British Machine Vision Association 2007 Sullivan Thesis Prize and was short listed for the British Computer Society 2007 Distinguished Dissertation Award. His research interests include visual search and object recognition applied to large image and video collections. After spending six months as a postdoctoral researcher in the Computer Science and Artificial Intelligence Laboratory at the Massachusetts Institute of Technology, he currently holds a permanent position as an INRIA researcher at the Departement d'Informatique, Ecole Normale Superieure, Paris. He has published over 50 scientific publications and serves as an Associate Editor of the International Journal of Computer Vision. He has been awarded an ERC Starting grant in 2013.

Masatoshi Okutomi Masatoshi Okutomi received a B.Eng. degree from the Department of Mathematical Engineering and Information Physics, the University of Tokyo, Japan, in 1981 and an M.Eng. degree from the Department of Control Engineering, Tokyo Institute of Technology, Japan, in 1983. He joined Canon Research Center, Canon Inc., Tokyo, Japan, in 1983. From 1987 to 1990, he was a visiting research scientist in the School of Computer Science at Carnegie Mellon University, USA. In 1993, he received a D.Eng. degree for his research on stereo vision from Tokyo Institute of Technology. Since 1994, he has been with Tokyo Institute of Technology, where he is currently a professor in the Department of Systems and Control Engineering, the School of Engineering.

Tomas Pajdla Tomas Pajdla received the MSc and PhD degrees from the Czech Technical University in Prague. He works in geometry and algebra of computer vision and robotics with emphasis on nonclassical cameras, 3D reconstruction, and industrial vision. He contributed to introducing epipolar geometry of panoramic cameras, noncentral camera models generated by linear mapping, generalized epipolar geometries, to developing solvers for minimal problems in structure from motion and to solving image matching problem. He coauthored works awarded prizes at OAGM 1998 and 2013, BMVC 2002 and ACCV 2014. He is a member of the IEEE. Google Scholar: http://scholar.google.com/citations?user=gnR4zf8AAAAJ