Image-based localization using LSTMs for structured feature correlation

F. Walch[1,3] C. Hazirbas[1] L. Leal-Taixé[1] T. Sattler[2] S. Hilsenbeck[3] D. Cremers[1]
[1]Technical University of Munich [2]Department of Computer Science, ETH Zürich [3]NavVis

Abstract

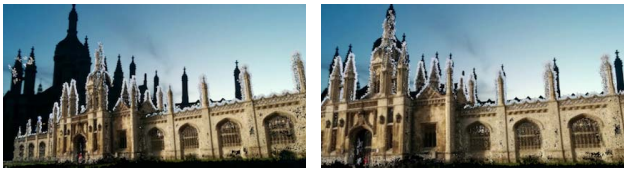
In this work we propose a new CNN+LSTM architecture for camera pose regression for indoor and outdoor scenes. CNNs allow us to learn suitable feature representations for localization that are robust against motion blur and illumination changes. We make use of LSTM units on the CNN output, which play the role of a structured dimensionality reduction on the feature vector, leading to drastic improvements in localization performance. We provide extensive quantitative comparison of CNN-based and SIFT-based localization methods, showing the weaknesses and strengths of each. Furthermore, we present a new large-scale indoor dataset with accurate ground truth from a laser scanner. Experimental results on both indoor and outdoor public datasets show our method outperforms existing deep architectures, and can localize images in hard conditions, e.g., *in the presence of mostly textureless surfaces, where classic SIFT-based methods fail.*

1. Introduction

Being able to localize a vehicle or device by estimating a camera pose from an image is a fundamental requirement for many computer vision applications such as navigating autonomous vehicles [34], mobile robotics and Augmented Reality [36], and Structure-from-Motion (SfM) [46].

Most state-of-the-art approaches [31, 44, 50, 62] rely on local features such as SIFT [35] to solve the problem of image-based localization. Given a SfM model of a scene, where each 3D point is associated with the image features from which it was triangulated, one proceeds in two stages [32, 43]: (i) establishing 2D-3D matches between features extracted from the query image and 3D points in the SfM model via descriptor matching; (ii) using these correspondences to determine the camera pose, usually by employing a n-point solver [29] inside a RANSAC loop [13]. Pose estimation can only succeed if enough correct matches have been found in the first stage. Consequently, limitations of both the feature detector, *e.g.*, motion blur or strong illumination changes, or the descriptor, *e.g.*, due to strong viewpoint changes, will cause localization approaches to fail.

Recently, two approaches have tackled the problem of localization with end-to-end learning. PlaNet [57] formulates localization as a classification problem, where the current position is matched to the best position in the training set. While this approach is suitable for localization in extremely large environments, it only allows to recover position but not orientation and its accuracy is bounded by the spatial extent of the training samples. More similar in spirit to our approach, PoseNet [23, 24] formulates 6DoF pose estimation as a regression problem. In this paper, we show that PoseNet is significantly less accurate than state-of-the-art SIFT methods [31, 44, 50, 62] and propose a novel network architecture that significantly outperforms PoseNet.



(a) PoseNet result [24] (b) Our result

Figure 1: Our approach achieves accurate outdoor image-based localization even in challenging lighting conditions where other deep architectures fail.

1.1. Contributions

In this paper, we propose to directly regress the camera pose from an input image. To do so, we leverage, on the one hand, Convolutional Neural Networks (CNNs) which allow us to learn suitable feature representations for localization that are more robust against motion blur and illumination changes. As we can see from the PoseNet [24] results, regressing the pose after the high dimensional output of a FC layer is not optimal. Our intuition is that the high dimensionality of the FC output makes the network prone to overfitting to training data. PoseNet deals with this problem with careful dropout strategies. We propose to make use of Long-Short Term Memory (LSTM) units [22] on the FC output, which performs structured dimensionality reduction and chooses the most useful feature correlations for the task of pose estimation. Overall, we improve localization accuracy by 32-37% wrt. previous deep learning architectures [23, 24]. Furthermore, we are the first to pro-

2380-7504/17 $31.00 © 2017 IEEE
DOI 10.1109/ICCV.2017.75

vide an extensive comparison with a state-of-the-art SIFT-based method [44], which shreds a light on the strengths and weaknesses of each approach. Finally, we introduce a new dataset for large-scale indoor localization, consisting of 1,095 high resolution images covering a total area of 5,575m^2. Each image is associated with ground truth pose information. We show that this sequence cannot be handled with SIFT-based methods, as it contains large textureless areas and repetitive structures. In contrast, our approach robustly handles this scenario and localizes images on average within 1.31m of their ground truth location.

To summarize, our contribution is three-fold: (i) we propose a new CNN+LSTM architecture for camera pose regression in indoor and outdoor scenes. Our approach significantly outperforms previous work on CNN-based localization [23, 24]. (ii) we provide the first extensive quantitative comparison of CNN-based and SIFT-based localization methods. We show that classic SIFT-based methods still outperform all CNN-based methods by a large margin on existing benchmark datasets. (iii) we introduce TUM-LSI[1], a new challenging large indoor dataset exhibiting repetitive structures and weakly textured surfaces, and provide accurate ground truth poses. We show that CNN-based methods can handle such a challenging scenario while SIFT-based methods fail completely. Thus, we are the first to demonstrate the usefulness of CNN-based methods in practice.

1.2. Related work

Local feature-based localization. There are two traditional ways to approach the localization problem. Location recognition methods represent a scene by a database of geo-tagged photos. Given a query image, they employ image retrieval techniques to identify the database photo most similar to the query [2, 42, 52, 53, 61]. The geo-tag of the retrieved image is often used to approximate the camera pose of the query, even though a more accurate estimate can be obtain by retrieving multiple relevant images [45, 60, 63].

More relevant to our approach are structure-based localization techniques that use a 3D model, usually obtained from Structure-from-Motion, to represent a scene. They determine the full 6DoF camera pose of a query photo from a set of 2D-3D correspondences established via matching features found in the query against descriptors associated with the 3D points. The computational complexity of matching grows with the size of the model. Thus, prioritized search approaches [10, 32, 44] terminate correspondence search as soon as a fixed number of matches has been found. Similarly, descriptor matching can be accelerated by using only a subset of all 3D points [9, 32], which at the same time reduces the memory footprint of the 3D models. The latter can also be achieved by quantizing the descriptors [36, 41].

[1]Dataset available at https://tum-lsi.vision.cs.tum.edu

For more complex scenes, *e.g.*, large-scale urban environments or even large collections of landmark scenes, 2D-3D matches are usually less unique as there often are multiple 3D points with similar local appearance [31]. This causes problems for the pose estimation stage as accepting more matches leads to more wrong matches and RANSAC's run-time grows exponentially with the ratio of wrong matches. Consequently, Sattler *et al*. use co-visibility information between 3D points to filter out wrong matches before pose estimation [41, 44]. Similarly, Li *et al*. use co-visibility information to adapt RANSAC's sampling strategy, enabling them to avoid drawing samples unlikely to lead to a correct pose estimate [31]. Assuming that the gravity direction and a rough prior on the camera's height are known, Svärm *et al*. propose an outlier filtering step whose run-time does not depend on the inlier ratio [50]. Zeisl *et al*. adapt this approach into a voting scheme, reducing the computational complexity of outlier filtering from $\mathcal{O}(n^2 \log n)$ [50] to $\mathcal{O}(n)$ for n matches [62].

The overall run-time of classical localization approaches depends on the number of features found in a query image, the number of 3D points in the model, and the number of found correspondences and/or the percentage of correct matches. In contrast, our approach directly regresses the camera pose from a single feed-forward pass through a network. As such, the run-time of our approach only depends on the size of the network used.

As we will show, SIFT-based methods do not work for our new challenging indoor LSI dataset due to repetitive structures and large textureless regions present in indoor scenes. This further motivates the use alternative approaches based, *e.g.*, on deep learning.

Localization utilizing machine learning. In order to boost location recognition performance, Gronat *et al*. and Cao & Snavely learn linear classifiers on top of a standard bag-of-words vectors [8, 18]. They divide the database into distinct places and train classifiers to distinguish between them.

Donoser & Schmalstieg cast feature matching as a classification problem, where the descriptors associated with each 3D model point form a single class [12]. They employ an ensemble of random ferns to efficiently compute matches.

Aubry *et al*. learn feature descriptors specifically for the task of localizing paintings against 3D scene models [3].

In the context of re-localization for RGB-D images, Guzman-Rivera *et al*. and Shotton *et al*. learn random forests that predict a 3D point position for each pixel in an image [19, 47]. The resulting 2D-3D matches are then used to estimate the camera pose using RANSAC. Rather than predicting point correspondences, Valentin *et al*. explicitly model the uncertainty of the predicted 3D point positions and use this uncertainty during pose estimation [54], allowing them to localize more images. Brachmann *et al*. adapt the random forest-based approach to not rely on depth mea-

surements during test time [6]. Still, they require depth data during the training stage as to predict 3D coordinates for each pixel. In contrast, our approach directly regresses the camera pose from an RGB image, and thus only needs a set of image-pose pairs as input for training.

Deep learning. CNNs have been successfully applied to most tasks in computer vision since their major success in image classification [21, 28, 48] and object detection [14, 15, 39]. One of the major drawbacks of deep learning is its need for large datasets for training. A common approach used for many tasks is that to fine-tune deep architectures pre-trained on the seemingly unrelated task of image classification on ImageNet [40]. This has been successfuly applied, among others, to object detection [14], object segmentation [27, 37], semantic segmentation [20, 38], and depth and normal estimation [30]. Similarly, we take pre-trained networks, *e.g*. GoogLeNet [51], which can be seen as feature extractors and then fine-tune them for the task of camera pose regression.

LSTM [22] is a type of Recurrent Neural Network (RNN) [16] designed to accumulate or forget relevant contextual information in its hidden state. It has been successfully applied for handwriting recognition [17] and in natural language processing for machine translation [49]. Recently, CNN and LSTM have been combined in the computer vision community to tackle, for example, visual recognition in videos [11]. While most methods in the literature apply LSTM on a temporal sequence, recent works have started to use the memory capabilities of LSTMs to encode contextual information. ReNet [56] replaced convolutions by RNNs sweeping the image vertically and horizontally. [55] uses spatial LSTM for person re-identification, parsing the detection bounding box horizontally and vertically in order to capture spatial dependencies between body parts. [7] employed the same idea for semantic segmentation and [33] for semantic object parsing. We use LSTMs to better correlate features coming out of the convolutional and FC layers, efficiently reducing feature dimensionality in a structured way that improves pose estimation compared to using dropout on the feature vector to prevent overfitting. A similar approach was simultaneously proposed in [4], where LSTMs are used to obtain contextual information for object recognition.

End-to-end learning has also been used for localization and location recognition. DSAC [5] proposes a differentiable RANSAC so that a matching function that optimizes pose quality can be learned. Arandjelović *et al*. employ CNNs to learn compact image representations, where each image in a database is represented by a single descriptor [1]. Weyand *et al*. cast localization as a classification problem [57]. They adaptively subdivide the earth's surface in a set of tiles, where a finer quantization is used for regions exhibiting more images. The CNN then learns to predict the corresponding tile for a given image, thus providing the approximate position from which a photo was taken. Focusing on accurate 6DoF camera pose estimation, the PoseNet method by Kendall *et al*. uses CNNs to model pose estimation as a regression problem [24]. An extension of the approach repeatedly evaluates the CNN with a fraction of its neurons randomly disabled, resulting in multiple different pose estimates that can be used to predict pose uncertainty [23]. One drawback of the PoseNet approach is its relative inaccuracy [6]. In this paper, we show how a CNN+LSTM architecture is able to produce significantly more accurate camera poses compared to PoseNet.

2. Deep camera pose regression

In this section, we develop our framework for learning to regress camera poses directly from images. Our goal is to train a CNN+LSTM network to learn a mapping from an image to a pose, $I \xrightarrow{f} \mathbf{P}$, where $f(\cdot)$ is the neural network. Each pose $\mathbf{P} = [\mathbf{p}, \mathbf{q}]$ is represented by its 3D camera position $\mathbf{p} \in \mathbb{R}^3$ and a quaternion $\mathbf{q} \in \mathbb{R}^4$ for its orientation.

Given a dataset composed of training images I_i and their corresponding 3D ground truth poses $\mathbf{P}_i$, we train the network using Adam [25] with the same loss function already used in [24]:

$$L_i = \|\mathbf{p}_i - \hat{\mathbf{p}}_i\|_2 + \beta \cdot \left\| \mathbf{q}_i - \frac{\hat{\mathbf{q}}_i}{\|\hat{\mathbf{q}}_i\|} \right\|_2 , \qquad (1)$$

where $(\mathbf{p}, \mathbf{q})$ and $(\hat{\mathbf{p}}, \hat{\mathbf{q}})$ are ground truth and estimated position-orientation pairs, respectively. We represent the orientation with quaternions and thus normalize the predicted orientation $\hat{\mathbf{q}}_i$ to unit length. β determines the relative weight of the orientation error wrt. to the positional error and is in general bigger for outdoor scenes, as errors tend to be relatively larger [24]. All hyperparameters used for the experiments are detailed in Section 4.

2.1. CNN architecture: feature extraction

Training a neural network from scratch for the task of pose regression would be impractical for several reasons: (i) we would need a really large training set, (ii) compared to classification problems, where each output label is covered by at least one training sample, the output in regression is continuous and infinite. Therefore, we leverage a pre-trained classification network, namely GoogLeNet [51], and modify it in a similar fashion as in [24]. At the end of the convolutional layers average pooling is performed, followed by a fully connected layer which outputs a 2048 dimensional vector (*c.f*. Figure 2). This can be seen as a feature vector that represents the image to be localized. This architecture is used in [24] to predict camera poses by using yet another fully connected regression layer at the end that outputs the 7-dimensional pose and orientation vector (the quaternion vector is normalized to unit length at test time).

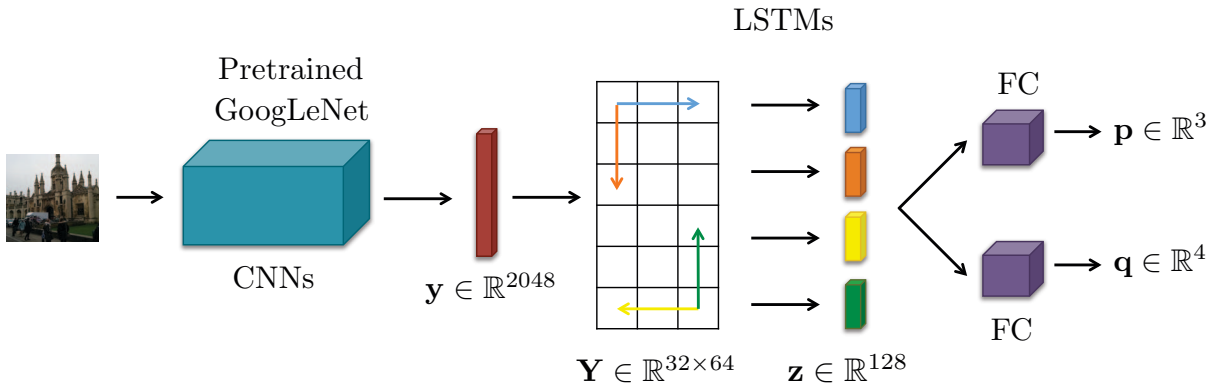


Figure 2: Architecture of the proposed pose regression LSTM network.

2.2. Structured feature correlation with LSTMs

After the convolutional layers of GoogleNet, an average pooling layer gathers the information of each feature channel for the entire image. Following PoseNet [24], we use a fully connected (FC) layer after pooling to learn the correlation among features. As we can see from the PoseNet [24] results shown in Section 4, regressing the pose after the high dimensional output of a fully connected (FC) layer is not optimal. Intuitively, the dimensionality of the 2048D embedding of the image through the FC layer is typically relatively large compared to the amount of available training data. As a result, the linear pose regressor has many degrees of freedom and it is likely that overfitting leads to inaccurate predictions for test images dissimilar to the training images. One could directly reduce the dimensions of the FC, but we empirically found that dimensionality reduction performed by a network with LSTM memory blocks is more effective. Compared to applying dropout within PoseNet to avoid overfitting [23], our approach consistently estimates more accurate positions, which justifies our use of LSTMs.

Even though Long Short-Term Memory (LTSM) units have been typically applied to temporal sequences, recent works [4, 7, 33, 55, 56] have used the memory capabilities of LSTMs in image space. In our case, we treat the output 2048 feature vector as our sequence. We propose to insert four LSTM units after the FC, which have the function of reducing the dimensionality of the feature vector in a structured way. The memory units identify the most useful feature correlations for the task of pose estimation.

Reshaping the input vector. In practice, inputting the 2048-D vector directly to the LSTM did not show good results. Intuitively, this is because even though the memory unit of the LSTM is capable of remembering distant features, a 2048 length vector is too long for LSTM to correlate from the first to the last feature. We thereby propose to reshape the vector to a 32×64 matrix and to apply four LSTMs in the up, down, left and right directions as depicted in Figure 2. These four outputs are then concatenated and passed to the fully connected pose prediction layers. This imitates the function of structured dimensionality reduction which greatly improves pose estimation accuracy.

3. A large-scale indoor localization dataset

Machine learning and in particular deep learning are inherently data-intensive endeavors. Specifically, supervised learning requires not only data but also associated ground truth labelling. For some tasks such as image classification [40] or outdoor image-based localization [24] large training and testing datasets have already been made available to the community. For indoor scenes, only small datasets covering a spatial extent the size of a room [47] are currently available.

We introduce the *TU Munich Large-Scale Indoor* (TUM-LSI) dataset covering an area that is two orders of magnitude larger than the typically used 7Scenes dataset [47]. It comprises 1,095 high-resolution images (4592×3448 pixels) with geo-referenced pose information for each image. The dataset spans a whole building floor with a total area of 5,575 m^2. Image locations are spaced roughly one meter apart, and at location each we provide a set of six wide-angle pictures, taken in five different horizontal directions (full 360°) and one pointing up (see Figure 3). Our new dataset is very challenging due to repeated structural elements with nearly identical appearance, e.g. two nearly identical stair cases, that create global ambiguities. Notice that such global ambiguities often only appear at larger scale and are thus missing from the 7Scenes dataset. In addition, there is a general lack of well-textured regions, whereas 7Scenes mostly depict highly textured areas. Both problems make this dataset challenging to approaches that only considers (relatively) small image patches.

In order to generate ground truth pose information for each image, we captured the data using the *NavVis M3*[2] indoor mapping platform. This mobile system is equipped with six *Panasonic* 16-Megapixel system cameras and three

[2]www.navvis.com

Figure 3: Example images from our TUM-LSI dataset. At each capture-location, we provide a set of six high-resolution wide-angle pictures, taken in five different horizontal directions and one pointing up.

Hokuyo laser range finders. Employing SLAM, the platform is able to reconstruct the full trajectory with sub-centimeter accuracy.

4. Experimental results

We present results on several datasets, proving the efficacy of our method in outdoor scenes like Cambridge [24] and small-scale indoor scenes such as 7Scenes [47]. The two datasets are very different from each other: 7Scenes has a very high number of images in a very small spatial extent, hence, it is more suited for applications such as Augmented Reality, while Cambridge Landmarks has sparser coverage and larger spatial extent, the perfect scenario for image-based localization. In the experiments, we show that our method can be applied to both scenarios and delivers competitive results. We provide comparisons to previous CNN-based approaches, as well as a state-of-the-art SIFT-based localization method [44]. Furthermore, we provide results for our new TUM-LSI dataset. SIFT-based methods fail on TUM-LSI due to textureless surfaces and repetitive structures, while our method is able to localize images with an average accuracy of 1.31m for an area of 5,575 m^2.

Experimental setup. We initialize the GoogLeNet part of the network with the Places [64] weights and randomly initialize the remaining weights. All networks take images of size 224×224 pixel as input. We use random crops during training and central crops during testing. A mean image is computed separately for each training sequence and is subtracted from all images. All experiments are performed on an NVIDIA Titan X using TensorFlow with Adam [25] for optimization. Random shuffling is performed for each batch, and regularization is only applied to weights, not biases. For all sequences we use the following hyperparameters: batch size 75, regularization $\lambda = 2^{-4}$, auxiliary loss weights $\gamma = 0.3$, dropout probability 0.5, and the parameters for Adam: $\epsilon = 1$, $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The β of Eq. 1 balances the orientation and positional penalties. To ensure a fair comparison, for Cambridge Landmarks and 7Scenes, we take the same values as PoseNet [24]: for the indoor scenes β is between 120 to 750 and outdoor scenes between 250 to 2000. For TUM-LSI, we set $\beta = 1000$.

Comparison with state-of-the-art. We compare results to two CNN-based approaches: PoseNet [24] and Bayesian PoseNet [23]. On Cambridge Landmarks and 7Scenes, results for the two PoseNet variants [23, 24] were taken directly from the author's publication [23]. For the new TUM-LSI dataset, their model was fine-tuned with the training images. The hyperparameters used are the same as for our method, except for the Adam parameter $\epsilon = 0.1$, which showed better convergence.

To the best of our knowledge, CNN-based approaches have not been quantitatively compared to SIFT-based localization approaches. We feel this comparison is extremely important to know how deep learning can make an impact in image-based localization, and what challenges are there to overcome. We therefore present results of a state-of-the-art SIFT-based method, namely Active Search [44].

Active Search estimates the camera poses wrt. a SfM model, where each 3D point is associated with SIFT descriptors extracted from the training images. Since none of the datasets provides both an SfM model and the SIFT descriptors, we constructed such models from scratch using VisualSFM [58,59] and COLMAP [46] and registered them against the ground truth poses of the training images. Thus, the camera poses reported for Active Search contain both the errors made by Active Search and the reconstruction and registration processes. The models used for localization do not contain any contribution from the testing images.

Active Search uses a visual vocabulary to accelerate descriptor matching. We trained a vocabulary containing 10k words from training images of the Cambridge dataset and a vocabulary containing 1k words from training images of the smaller 7Scenes dataset. Active Search uses these vocabularies for prioritized search for efficient localization, where matching is terminated once a fixed number of correspondences has been found. We report results both with (w/) and without (w/o) prioritization. In the latter case, we simply do not terminate matching early but try to find as many corre-

Table 1: Median localization results of several RGB-only methods on Cambridge Landmarks [24] and 7Scenes [47].

Scene	Area or Volume	Active Search (w/o) [44]	Active Search (w/) [44]	PoseNet [24]	Bayesian PoseNet [23]	Proposed + Improvement(pos,ori)
King's College	5600 m^2	0.42 m, 0.55° (0)	0.57 m, 0.70° (0)	1.92 m, 5.40°	1.74 m, 4.06°	0.99 m, 3.65° (48,32)
Old Hospital	2000 m^2	0.44 m, 1.01° (2)	0.52 m, 1.12° (2)	2.31 m, 5.38°	2.57 m, 5.14°	1.51 m, 4.29° (35,20)
Shop Façade	875 m^2	0.12 m, 0.40° (0)	0.12 m, 0.41° (0)	1.46 m, 8.08°	1.25 m, 7.54°	1.18 m, 7.44° (19,8)
St Mary's Church	4800 m^2	0.19 m, 0.54° (0)	0.22 m, 0.62° (0)	2.65 m, 8.48°	2.11 m, 8.38°	1.52 m, 6.68° (43,21)
Average All	–	–	–	2.08 m, 6.83°	1.92 m, 6.28°	1.30 m, 5.52° (37,19)
Average by [44]	–	0.29 m, 0.63°	0.36 m, 0.71°	–	–	1.37 m, 5.52°
Chess	6 m^3	0.04 m, 1.96° (0)	0.04 m, 2.02° (0)	0.32 m, 8.12°	0.37 m, 7.24°	0.24 m, 5.77° (25,29)
Fire	2.5 m^3	0.03 m, 1.53° (1)	0.03 m, 1.50° (1)	0.47 m, 14.4°	0.43 m, 13.7°	0.34 m, 11.9° (28,17)
Heads	1 m^3	0.02 m, 1.45° (1)	0.02 m, 1.50° (1)	0.29 m, 12.0°	0.31 m, 12.0°	0.21 m, 13.7° (27,-14)
Office	7.5 m^3	0.09 m, 3.61° (34)	0.10 m, 3.80° (34)	0.48 m, 7.68°	0.48 m, 8.04°	0.30 m, 8.08° (37,-5)
Pumpkin	5 m^3	0.08 m, 3.10° (71)	0.09 m, 3.21° (68)	0.47 m, 8.42°	0.61 m, 7.08°	0.33 m, 7.00° (30,17)
Red Kitchen	18 m^3	0.07 m, 3.37° (0)	0.07 m, 3.52° (0)	0.59 m, 8.64°	0.58 m, 7.54°	0.37 m, 8.83° (37,-2)
Stairs	7.5 m^3	0.03 m, 2.22° (3)	0.04 m, 2.22° (0)	0.47 m, 13.8°	0.48 m, 13.1°	0.40 m, 13.7° (15,0.7)
Average All	–	–	–	0.44 m, 10.4°	0.47 m, 9.81°	0.31 m, 9.85° (29,5)
Average by [44]	–	0.05 m, 2.46°	0.06 m, 2.54°	–	–	0.30 m, 9.15°

spondences as possible. For querying with Active Search, we use calibrated cameras with a known focal length, obtained from the SfM reconstructions, but ignore radial distortion. As such, camera poses are estimated using a 3-point-pose solver [26] inside a RANSAC loop [13]. Poses estimated from only few matches are usually rather inaccurate. Following common practice [31, 32], Active Search only considers a testing image as successfully localized if its pose was estimated from at least 12 inliers.

4.1. Large-scale outdoor localization

We present results for outdoor image-based localization on the publicly available Cambridge Landmarks dataset [24] in Table 1. We report results for Active Search only for images with at least 12 inliers and give the number of images where localization fails in parenthesis. In order to compare the methods fairly, we provide the average accuracy for all images (Average All), and also the average accuracy for only those images that [44] was able to localize (Average by [44]). Note, that we do not report results on the Street dataset of Cambridge Landmarks. It is a unique sequence because the training database consists of four distinct video sequences, each filmed in a different compass direction. This results in training images at similar positions, but with very different orientations. Even with the hyperparameters set by the author of [24], training did not converge for any of the implemented methods.

Figure 4: The class activation map is overlaid on an input image from King's College as a heat map. Red areas indicate parts of the image the network considers important for pose regression. The visualization shows how the network focuses on distinctive building elements.

In parenthesis next to our results, we show the rounded percentage improvement wrt. PoseNet in both position and orientation, separated by a comma. As we can see, the proposed method on average reduces positional error by 37.5% wrt. PoseNet and the orientation error by 19%. For example, in King's College the positional error is reduced by more than 40%, going from 1.92m for Posenet to 0.99m for our method. This shows that the proposed LSTM-based structured output is efficient in encoding feature correlations, leading to large improvements in localization performance. It is important to note that none of the CNN-based methods is able to match the precision of Active Search [44], especially when computing the orientation of the camera. Since [44] requires 12 inliers to consider an image as localized, it is able to reject inaccurate poses. In contrast, our method always provides a localization result, even if it is sometimes less accurate. Depending on the application, one behavior or the other might be more desirable. As an example, we show in Figure 6 an image from Old Hospital, where a tree is occluding part of the building. In this case, [44] is not able to localize the image, while our method still produces a reasonably accurate pose. This phenomenon becomes more important in indoor scenes, where Active Search is unable to localize a substan-

tially larger number of images, mostly due to motion blur.

Interestingly, for our method, the average for all images is lower than the average only for the images that [44] can also localize. This means that for images where [44] cannot return a pose, our method actually provides a very accurate result. This "complementary" behavior between SIFT- and CNN-based methods could be exploitable in future research. Overall, our method shows a strong performance in outdoor image-based localization, as seen in Figure 7, where PoseNet [24] provides less accurate poses. In order to better understand how the network localizes an image, Figure 4 plots the class activation maps for the King's College sequence. Notice how strong activations cluster around distinctive building elements, *e.g.*, the towers and entrance.

4.2. Small-scale indoor localization

In this section, we focus on localization on small indoor spaces, for which we use the publicly available 7Scenes dataset [47]. Results at the bottom of Table 1 show that we also outperform previous CNN-based PoseNet by 29% in positional error and 5.3% orientation error. For example on, Pumpkin we achieve a positional error reduction from the 0.61m for Posenet to 0.33m for our method. As for the Cambridge Landmarks dataset, we observe that our approach consistently outperforms Bayesian PoseNet [23], which uses dropout to limit overfitting during training. These experimental results validate our strategy of using LSTMs for structured dimensionality reduction in an effort to avoid overfitting.

There are two methods that use RGB-D data and achieve a lower error but still higher than Active Search: [6] achieves 0.06m positional error and 2.7°orientation error, while [47] scores 0.08m and 1.60°. Note, that these methods require RGB-D data for training [6] and/or testing [47]. It is unclear though how well such methods would work in outdoor scenarios with stereo data. In theory, multi-view stereo methods could be used to obtain the required depth maps for outdoor scenes. However, such depth maps are usually substantially more noisy and contain significantly more outliers compared to the data obtained with RGB-D sensors. In addition, the accuracy of the depth maps decreases quadratically with the distance to the scene, which is usually much larger for outdoor scenes than for indoor scenes. In [54], authors report that 63.4% of all test images for Stairs can be localized with a position error less than 5cm and an orientation error smaller than 5°. With and without prioritization, Active Search localizes 77.8% and 80.2%, respectively, within these error bounds. Unfortunately, median registration errors of 2-5cm observed when registering the other SfM models against the 7Scene datasets prevent us from a more detailed comparison.

As we can see in Table 1, if an image can be localized, we notice that Active Search performs better than CNN-based approaches. However, we note that for Office and Pumpkin the number of images not localized is fairly large, 34 and 71, respectively. We provide the average accuracy for all images (Average All), and also the average accuracy for only those images that [44] was able to localize (Average by [44]). Note that for our method, the two averages are extremely similar, *i.e.*, we are able to localize those images with the same accuracy as all the rest, showing robustness wrt. motion blur that heavily affect SIFT-based methods. This shows the potential of CNN-based methods.

Table 2: Median localization accuracy on TUM-LSI.

Area	# train/test	PoseNet [24]	Proposed
5575 m^2	875/220	1.87 m, 6.14°	1.31 m, 2.79° (30,55)

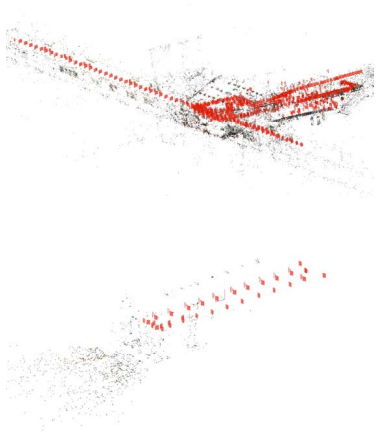


Figure 5: Failed SfM reconstructions for the TUM-LSI dataset, obtained with COLMAP. The first reconstruction contains two stairwells collapsed into one due to repetitive structures. Due to a lack of sufficient matches, the method was unable to connect a sequence of images and therefore creates a second separate model of one of the hallways.

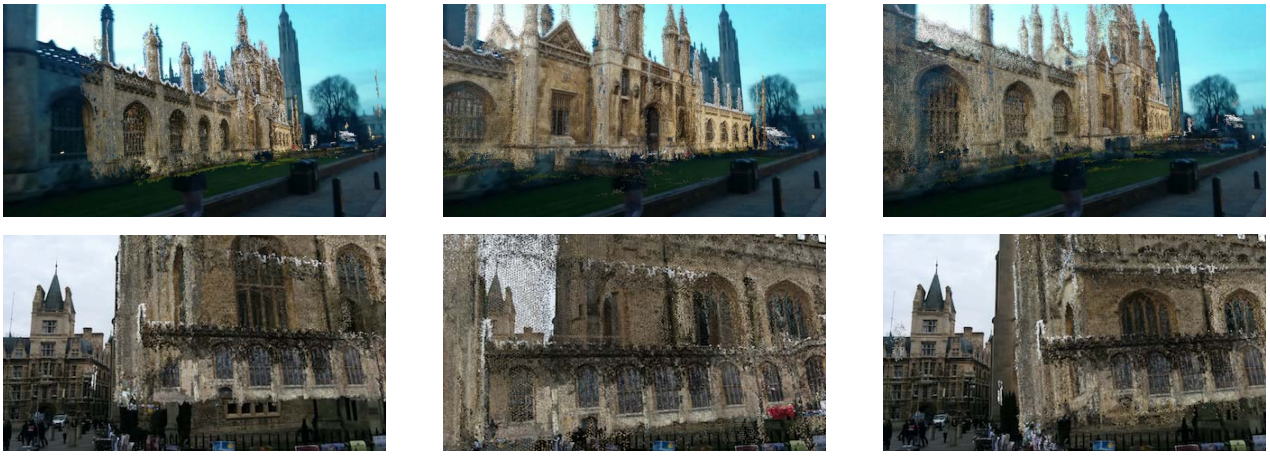
4.3. Complex large-scale indoor localization

In our last experiment, we present results on the new TUM-LSI dataset. It covers a total area of 5,575 m^2, the same order of magnitude as the outdoor localization dataset, and much larger than typical indoor datasets like 7Scenes.

Figure 3 shows an example from the dataset that contains large textureless surfaces. These surfaces are known to cause problems for methods based on local features. In fact, we were not able to obtain correct SfM reconstructions for the TUM-LSI dataset. The lack of texture in most parts of the images, combined with repetitive scene elements, causes both VisualSFM and COLMAP to fold repetitive structures onto themselves. For example, the two separate stairwells (red floor in Figure 3) are mistaken for a single stairwell (*c.f.* Figure 5). As the resulting 3D model

(a) Original image (b) Posenet result [24] (c) Our result

Figure 6: Examples from the Old Hospital sequence. The 3D scene model is projected into the image using the poses estimated by (b) PoseNet [24] and (c) our method. Active Search [44] did not localize the image due to the occlusion caused by the tree. Note the inaccuracy of PoseNet compared to the proposed method (check the top of the building for alignment).



(a) Active Search (w/) [44] result (b) Posenet result [24] (c) Our result

Figure 7: Examples of localization results on King's College for Active Search [44], PoseNet [24], and the proposed method.

does not reflect the true 3D structure of the scene, there is no point in applying Active Search or any other SIFT-based method. Notice that such repetitive structures would cause Active Search to fail even if a good model is provided.

For the experiments on the TUM-LSI dataset, we ignore the ceiling-facing cameras. As we can see in Table 2, our method outperforms PoseNet [24] by almost 30% in positional error and 55% orientation error, showing a similar improvement as for other datasets. To the best of our knowledge, we are the first to showcase a scenario where CNN-based methods succeed while SIFT-based approaches fail. On this challenging sequence, our method achieves an average error of around 1m. In our opinion, this demonstrates that CNN-based methods are indeed a promising avenue to tackle hard localization problems such as repetitive structures and textureless walls, which are predominant in modern buildings, and are a problem for classic SIFT-based localization methods.

5. Conclusion

In this paper, we address the challenge of image-based localization of a camera or an autonomous system with a novel deep learning architecture that combines a CNN with LSTM units. Rather than precomputing feature points and building a map as done in traditional SIFT-based localization techniques, we determine a direct mapping from input image to camera pose. With a systematic evaluation on existing indoor and outdoor datasets, we show that our LSTM-based structured feature correlation can lead to drastic improvements in localization performance compared to other CNN-based methods. Furthermore, we are the first to show a comparison of SIFT-based and CNN-based localization methods, showing that classic SIFT approaches still outperform all published CNN-based methods to date on standard benchmark datasets. To answer the ensuing question whether CNN-based localization is a promising direction of research, we demonstrate that our approach succeeds in a very challenging scenario where SIFT-based methods fail. To this end, we introduce a new challenging large-scale indoor sequence with accurate ground truth. Besides aiming to close the gap in accuracy between SIFT- and CNN-based methods, we believe that exploring CNN-based localization in hard scenarios is a promising research direction.

Acknowledgements. This work was partially funded by the ERC Consolidator grant *3D Reloaded* and a Sofja Kovalevskaja Award from the Alexander von Humboldt Foundation, endowed by the Federal Ministry of Education and Research.

References

[1] R. Arandjelović, P. Gronat, A. Torii, T. Pajdla, and J. Sivic. NetVLAD: CNN architecture for weakly supervised place recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 3

[2] R. Arandjelović and A. Zisserman. DisLocation: Scalable descriptor distinctiveness for location recognition . In *Asian Conference on Computer Vision (ACCV)*, 2014. 2

[3] Aubry, Mathieu and Russell, Bryan C. and Sivic, Josef. Painting-to-3D Model Alignment via Discriminative Visual Elements. *ACM Transactions on Graphics (TOG)*, 2014. 2

[4] S. Bell, C. L. Zitnick, K. Bala, and R. Girshick. Inside-Outside Net: Detecting Objects in Context with Skip Pooling and Recurrent Neural Networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 3, 4

[5] E. Brachmann, A. Krull, S. Nowozin, J. Shotton, F. Michel, S. Gumhold, and C. Rother. DSAC - Differentiable RANSAC for Camera Localization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 3

[6] E. Brachmann, F. Michel, A. Krull, M. Y. Yang, S. Gumhold, and C. Rother. Uncertainty-driven 6d pose estimation of objects and scenes from a single rgb image. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 3, 7

[7] W. Byeon, T. M. Breuel, F. Raue, and M. Liwicki. Scene labeling with lstm recurrent neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 3, 4

[8] S. Cao and N. Snavely. Graph-Based Discriminative Learning for Location Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013. 2

[9] S. Cao and N. Snavely. Minimal Scene Descriptions from Structure from Motion Models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 2

[10] S. Choudhary and P. J. Narayanan. Visibility probability structure from sfm datasets and applications. In *European Conference on Computer Vision (ECCV)*, 2012. 2

[11] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 3

[12] M. Donoser and D. Schmalstieg. Discriminative Feature-to-Point Matching in Image-Based Locallization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 2

[13] M. Fischler and R. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM (CACM)*, 1981. 1, 6

[14] R. Girshick. Fast R-CNN. In *IEEE International Conference on Computer Vision (ICCV)*, 2015. 3

[15] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurateobject detection and semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 3

[16] C. Goller and A. Kuchler. Learning task-dependent distributed representations by backpropagation through structure. In *IEEE International Conference on Neural Networks (ICNN)*, 1996. 3

[17] A. Graves and J. Schmidhuber. Offline handwriting recognition with multidimensional recurrent neural networks. In *Conference on Neural Information Processing Systems (NIPS)*, 2009. 3

[18] P. Gronat, G. Obozinski, J. Sivic, and T. Pajdla. Learning per-location classifiers for visual place recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013. 2

[19] A. Guzman-Rivera, P. Kohli, B. Glocker, J. Shotton, T. Sharp, A. Fitzgibbon, and S. Izadi. Multi-Output Learning for Camera Relocalization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 2

[20] C. Hazirbas, L. Ma, C. Domokos, and D. Cremers. Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture. In *Asian Conference on Computer Vision (ACCV)*, 2016. 3

[21] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 3

[22] S. Hochreiter and J. Schmidhuber. Long short-term memory. In *Neural Computation (NECO)*, 1997. 1, 3

[23] A. Kendall and R. Cipolla. Modelling uncertainty in deep learning for camera relocalization. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2016. 1, 2, 3, 4, 5, 6, 7

[24] A. Kendall, M. Grimes, and R. Cipolla. Posenet: A convolutional network for real-time 6-dof camera relocalization. In *IEEE International Conference on Computer Vision (ICCV)*, 2015. 1, 2, 3, 4, 5, 6, 7, 8

[25] D. P. Kingma and J. L. Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations (ICLR)*, 2015. 3, 5

[26] L. Kneip, D. Scaramuzza, and R. Siegwart. A novel parametrization of the perspective-three-point problem for a direct computation of absolute camera position and orientation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011. 6

[27] I. Kokkinos. Pushing the boundaries of boundary detection using deep learning. In *International Conference on Learning Representations (ICLR)*, 2016. 3

[28] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In *Conference on Neural Information Processing Systems (NIPS)*, 2012. 3

[29] Z. Kukelova, M. Bujnak, and T. Pajdla. Real-time solution to the absolute pose problem with unknown radial distortion and focal length. In *IEEE International Conference on Computer Vision (ICCV)*, 2013. 1

[30] B. Li, C. Shen, Y. Dai, A. van den Hengel, and M. He. Depth and surface normal estimation from monocular images using regression on deep features and hierarchical crfs. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 3

[31] Y. Li, N. Snavely, D. Huttenlocher, and P. Fua. Worldwide pose estimation using 3d point clouds. In *European Conference on Computer Vision (ECCV)*, 2012. 1, 2, 6

[32] Y. Li, N. Snavely, and D. P. Huttenlocher. Location Recognition Using Prioritized Feature Matching. In *European Conference on Computer Vision (ECCV)*, 2010. 1, 2, 6

[33] X. Liang, X. Shen, D. Xiang, J. Feng, L. Lin, and S. Yan. Semantic object parsing with local-global long short-term memory. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 3, 4

[34] H. Lim, S. N. Sinha, M. F. Cohen, and M. Uyttendaele. Real-time image-based 6-dof localization in large-scale environments. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012. 1

[35] D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision (IJCV)*, 2004. 1

[36] S. Lynen, T. Sattler, M. Bosse, J. Hesch, M. Pollefeys, and R. Siegwart. Get Out of My Lab: Large-scale, Real-Time Visual-Inertial Localization. In *Robotics: Science and Systems (RSS)*, 2015. 1, 2

[37] K. Maninis, J. Pont-Tuset, P. Arbeláez, and L. V. Gool. Convolutional oriented boundaries. In *European Conference on Computer Vision (ECCV)*, 2016. 3

[38] H. Noh, S. Hong, and B. Han. Learning deconvolution network for semantic segmentation. *IEEE International Conference on Computer Vision (ICCV)*, 2015. 3

[39] S. Ren, K. H. andRoss Girshick, and J. Sun. Faster R-CNN: Towards real-time object detectionwith region proposal networks. In *Conference on Neural Information Processing Systems (NIPS)*, 2015. 3

[40] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 2015. 3, 4

[41] T. Sattler, M. Havlena, F. Radenović, K. Schindler, and M. Pollefeys. Hyperpoints and Fine Vocabularies for Large-Scale Location Recognition. In *IEEE International Conference on Computer Vision (ICCV)*, 2015. 2

[42] T. Sattler, M. Havlena, K. Schindler, and M. Pollefeys. Large-Scale Location Recognition And The Geometric Burstiness Problem. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2

[43] T. Sattler, B. Leibe, and L. Kobbelt. Fast image-based localization using direct 2d-to-3d matching. In *IEEE International Conference on Computer Vision (ICCV)*, 2011. 1

[44] T. Sattler, B. Leibe, and L. Kobbelt. Efficient & Effective Prioritized Matching for Large-Scale Image-Based Localization. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2016 (to appear). 1, 2, 5, 6, 7, 8

[45] T. Sattler, A. Torii, J. Sivic, M. Pollefeys, H. Taira, M. Okutomi, and T. Pajdla. Are Large-Scale 3D Models Really Necessary for Accurate Visual Localization? In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[46] J. L. Schönberger and J.-M. Frahm. Structure-from-motion revisited. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1, 5

[47] J. Shotton, B. Glocker, C. Zach, S. Izadi, A. Criminisi, and A. Fitzgibbon. Scene coordinate regression forests for camera relocalization in rgb-d images. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013. 2, 4, 5, 6, 7

[48] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *International Conference on Learning Representations (ICLR)*, 2015. 3

[49] I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. In *Conference on Neural Information Processing Systems (NIPS)*, 2014. 3

[50] L. Svärm, O. Enqvist, F. Kahl, and M. Oskarsson. City-Scale Localization for Cameras with Known Vertical Direction. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2016 (to appear). 1, 2

[51] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 3

[52] A. Torii, R. Arandjelović, J. Sivic, M. Okutomi, and T. Pajdla. 24/7 place recognition by view synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 2

[53] A. Torii, J. Sivic, T. Pajdla, and M. Okutomi. Visual Place Recognition with Repetitive Structures. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013. 2

[54] J. Valentin, M. Nießner, J. Shotton, A. Fitzgibbon, S. Izadi, and P. Torr. Exploiting Uncertainty in Regression Forests for Accurate Camera Relocalization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 2, 7

[55] R. R. Varior, B. Shuai, J. Lu, D. Xu, and G. Wang. A siamese long short-term memory architecture for human re-identification. In *European Conference on Computer Vision (ECCV)*, 2016. 3, 4

[56] F. Visin, K. Kastner, K. Cho, M. Matteucci, A. Courville, and Y. Bengio. Renet: A recurrent neural network based alternative to convolutional networks. *arXiv preprint arXiv:1505.00393*, 2015. 3, 4

[57] T. Weyand, I. Kostrikov, and J. Philbin. Planet - photo geolocation with convolutional neural networks. *European Conference on Computer Vision (ECCV)*, 2016. 1, 3

[58] C. Wu. Towards linear-time incremental structure from motion. In *International Conference on 3D Vision (3DV)*, 2013. 5

[59] C. Wu, S. Agarwal, B. Curless, and S. M. Seitz. Multicore Bundle Adjustment. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011. 5

[60] A. R. Zamir and M. Shah. Accurate image localization based on google maps street view. In *European Conference on Computer Vision (ECCV)*, 2010. 2

[61] A. R. Zamir and M. Shah. Image Geo-Localization Based on Multiple Nearest Neighbor Feature Matching Using Gen-

eralized Graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2014. 2

[62] B. Zeisl, T. Sattler, and M. Pollefeys. Camera pose voting for large-scale image-based localization. In *IEEE International Conference on Computer Vision (ICCV)*, 2015. 1, 2

[63] W. Zhang and J. Kosecka. Image based localization in urban environments. In *International Symposium on 3D Data Processing, Visualization, and Transmission (3DPVT)*, 2006. 2

[64] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning deep features for scene recognition using places database. In *Conference on Neural Information Processing Systems (NIPS)*, 2014. 5