Deep Regression for Monocular Camera-based 6-DoF Global Localization in Outdoor Environments

Tayyab Naseer and Wolfram Burgard

***Abstract*— Precise localization of robots is imperative for their safe and autonomous navigation in both indoor and outdoor environments. In outdoor scenarios, the environment typically undergoes significant perceptual changes and requires robust methods for accurate localization. Monocular camera-based approaches provide an inexpensive solution to such challenging problems compared to 3D LiDAR-based methods. Recently, approaches have leveraged deep convolutional neural networks (CNNs) to perform place recognition and they turn out to outperform traditional handcrafted features under challenging perceptual conditions. In this paper, we propose an approach for directly regressing a 6-DoF camera pose using CNNs and a single monocular RGB image. We leverage the idea of transfer learning for training our network as this technique has shown to perform better when the number of training samples are not very high. Furthermore, we propose novel data augmentation in 3D space for additional pose coverage which leads to more accurate localization. In contrast to the traditional visual metric localization approaches, our resulting map size is constant with respect to the database. During localization, our approach has a constant time complexity of $\mathcal{O}(1)$ and is independent of the database size and runs in real-time at$\sim$80 Hz using a single GPU. We show the localization accuracy of our approach on publicly available datasets and that it outperforms CNN-based state-of-the-art methods.**

I. INTRODUCTION

Robust monocular camera-based autonomous navigation in outdoor environments is still a challenging problem. Precise metric localization is of paramount importance for an autonomous platform. Monocular global localization can be broadly categorized into visual place recognition and 6-DoF camera pose estimation. The approaches which fall into the former category, recognize the same place when a robot revisits. These methods provide topological localization but do not provide exact camera position. Approaches from the latter category estimate camera's global position and orientation in the map. In this paper, we address the latter problem as we believe that inferring the location of the robot in the map is crucial for its safe and autonomous navigation.

Most of the approaches [10, 20, 25] rely on local feature descriptors such as SIFT by Lowe [12] to cope with the problem of image-based localization. For a given 3D model of an environment, each point is associated with its image features that are used for triangulation. These approaches then establish 2D-3D correspondences between the query descriptors and the descriptors of 3D points. The correspondences are used to estimate the camera pose with a Perspective-n-Point solver. This is prone to outliers in the set of point correspondences. Generally, these pose estimates are further refined with RANSAC to cope with the outliers. Pose estimation in such a framework highly relies on correct feature matches. Image degradations like blur, poor illumination, and perceptual changes affect the feature descriptions and hence lead to poor localization accuracy.

All the authors are with the Department of Computer Science, University of Freiburg, 79110, Germany.

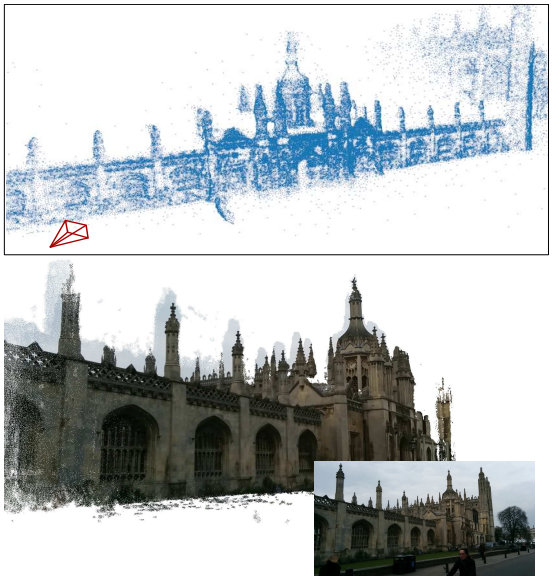

Fig. 1: Given an RGB image, our approach regresses the global 6-DoF camera pose in a large outdoor map. *Top:* 3D point cloud of the map (shown for visualization purpose). *Middle:* projection of the 3D cloud by the estimated camera pose using our approach. The projected point cloud is colorized with the color information of the test image. *Bottom:* Test image captured from a hand-held camera.

Recently, convolutional neural networks (CNNs) have shown tremendous progress in the area of visual place recognition in such perceptually challenging conditions [13]. Inspired by the amazing ability of such networks to perform well under harsh visual conditions, Kendall and Cipolla [7, 8] (Bayesian PoseNet, PoseNet) have explored the area of directly regressing the camera pose from these networks. In this paper, we propose to regress the 6-DoF camera pose from a single monocular RGB image as shown in Fig. 1. We leverage pre-trained CNNs on the Places database [26] for the regression task. This method is known as transfer learning. Kendall et al. [8] has recently shown that the networks that are pre-trained for classification tasks can be deployed to perform regression as well. Our idea in

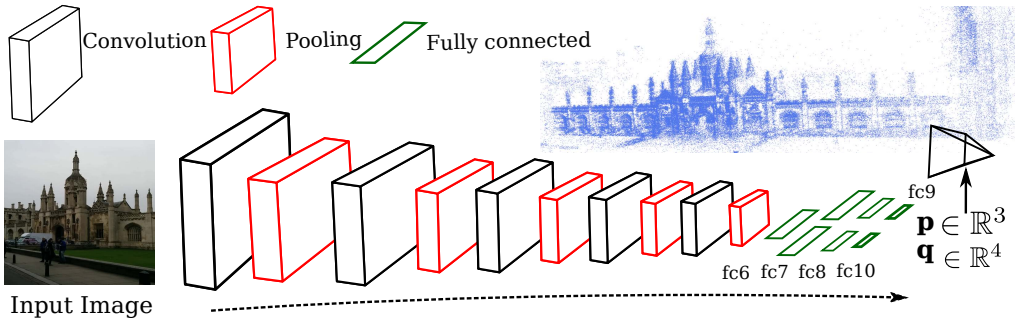

Fig. 2: Our proposed deep regression architecture with final pose regressors that predict position and orientation for an input image.

spirit is similar to PoseNet. We propose a novel CNN-based architecture for our regression task which builds upon the VGG architecture [21]. We segregate the fully connected layers for independent orientation and position prediction. Two fully connected layers (pose regressors) are added after the respective low dimensional features. The proposed architecture achieves better performance than vanilla VGG with pose regressors as the final branches. Furthermore, we propose data augmentation in the 3D pose space to generate more training examples which leads to better performance. We augment the images using a single monocular image. We leverage deep networks to generate depth maps from a single image. Together with the original RGB image and the corresponding depth map, we generate synthetic poses-image pairs. This helps us to expand the span of the pose space while training our network. To show that our method generalizes well to every scenario in the benchmark we do not perform grid search to optimize a main weighting parameter. We show with real-world experiments on the publicly available datasets that our approach outperforms CNN-based state-of-the-art methods.

Our contributions can be summarized as follows:

- We propose a segregated architecture for 6-DoF camera pose estimation.
- We propose data augmentation in 3D space for the task of camera relocalization. This helps in regressing the camera pose more accurately as the network is able to learn a more discriminative regression function.
- Our approach generalizes well to different scenarios of a public benchmark without performing a grid search to choose the main weighting parameter that balances the importance of orientation and position errors.

II. Related Work

Although visual localization has received great attention in computer vision and robotics communities, it still remains a challenging problem. We categorize these into topological localization and metric localization.

Topological Localization: Given a query image and a set of database images, these approaches retrieve the closest place in the map by leveraging different feature matching strategies. Milford and Wyeth [15] proposed an approach for place recognition across large perceptual changes by performing linear sequential filtering on image matchings. Sünderhauf et al. [22] proposed to leverage robustness of convolutional features with region proposals for accurate topological localization. Badino et al. [1] proposed an approach which fuses LiDAR and image data with a particle filter framework to perform longterm place recognition. Neubert and Protzel [17] propose a multi scale approach based on superpixel segmentation for robust place recognition. Although these approaches have shown impressive results in challenging conditions, these do not provide metric information about the 6-DoF pose of the camera. Torii et al. [23] used Google Street view images and corresponding depth maps to synthesize virtual views to boost the place recognition performance.

Metric Localization: McManus et al. [14] proposed an approach for learning salient visual elements of a place using a bank of SVM classifiers. This approach is hybrid as it uses weak localizers to find the closest topological node in the map and then refines the pose using the bank of SVM classifiers per place. It achieves sub-meter localization accuracy and requires 10 MB storage per place. Pascoe et al. [18] proposed an approach to localize camera images in a map built by fusing LIDAR and image data. Caselitz et al. [2] proposed to match geometry of images to the geometrical structure of a map built from 3D LiDAR data to cope with large perceptual changes. Visual SLAM, and vision-based localization approaches mostly focus on matching viewpoints using point-based features [3, 16, 19]. Kendall et al. [8] proposed to directly regress the camera pose from a monocular image in an end-to-end fashion. Kendall and Cipolla [7] showed that modeling the uncertainty in camera pose estimates can lead to better localization performance. A very recent method by Walch et al. [24] proposed to learn contextual features of images using spatial LSTMs [5] combined with PoseNet architecture to improve the localization accuracy. Our idea is similar in spirit to PoseNet with key differences in the architecture and the training strategies. We show the generalization of our approach which uses same training parameters in contrast to these recent CNN-based methods that use hyperparameter optimization for each dataset. In the next section, we discuss the technical contributions of our approach followed by extensive set of evaluations and show that our approach outperforms recent methods for 6-DoF camera pose regression in outdoor environments.

III. Technical Approach

Deep learning-based approaches have shown immense impact in the area of image classification and recognition. In this paper, we propose a method to achieve robust metric global monocular localization using convolutional neural networks in an end-to-end fashion. In this work, we build upon existing convolutional architectures and propose effective modifications for camera pose regression. We propose a novel CNN architecture which segregates the fully connected layers to estimate the position and orientation independently. Furthermore, we create synthetic viewpoints from the training images to prevent the network from over fitting on the datasets with small number of training examples. Such an augmentation in 3D space helps to learn a more discriminative regression function. In the following subsections, we discuss the proposed architecture and the training strategy.

A. Regression Conv-Net Architecture

In this subsection, we discuss the proposed convolutional neural network architecture for regression. We build upon the VGG16 architecture for directly regressing the 6-DoF pose from a single monocular RGB image in an end-to-end manner. It uses small receptive field of size 3×3 through out the network. It stacks several convolutional layers in conjunction to approximate a conv-layer of greater receptive field. This results in reduced trainable parameters. The proposed architecture is shown in Fig. 2. VGG16 has 3 fully connected layers ($\mathsf{fc6}, \mathsf{fc7}, \mathsf{fc8}$) after the convolutional layers. We branch out the network after the first fully connected layer to regress the camera position and orientation separately. We add two fully connected layers $\mathsf{fc9}$ and $\mathsf{fc10}$ at the end which are the final *pose regressors* for position and orientation. Dropout layers are added after each fully connected layers except the pose regressors to perform regularization. The pose regressors are initialized with Xavier weights as it prevents the input signal from shrinking or exploding based on the initial values of the weights [4]. For a fully connected layer, the variance of the weights $\mathcal{W}$ depends on its input and output dimensionality.

$$Var(\mathcal{W}) = \frac{2}{n_{in} + n_{out}} \tag{1}$$

All the remaining layers are initialized with the weights of VGG16 that is pre-trained on the Places database. We train our network with Adam [9] as the gradient descent solver.

B. Regression Conv-Net Training

Given a input image $\mathcal{I}$ and network parameters θ, our network predicts 6-DoF camera pose as two disjoint vectors. The output vector consists of the 3D camera position $\mathbf{p}$ and its orientation $\mathbf{q}$ represented as a quaternion.

Our aim is to optimize the following objective function which minimizes the euclidean loss between position and orientation estimate predictions and the true labels.

$$\mathcal{L}(\theta) = \|\hat{\mathbf{p}} - \mathbf{p}\|_2 + \beta \|\hat{\mathbf{q}} - \mathbf{q}\|_2 \tag{2}$$

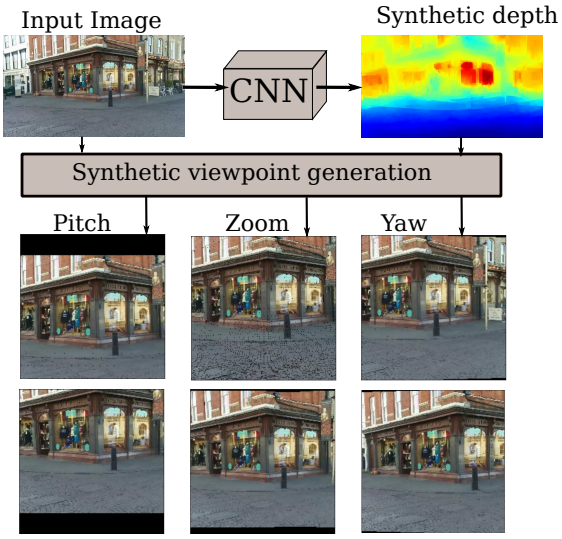

Fig. 3: We generate synthetic depth maps from a monocular RGB image using a Convnet. We then generate synthetic viewpoints using this depth information which help to learn a more discriminative regression function. We rotate the camera $\pm 5^\circ$ around pitch and yaw and translate by $\pm$ 0.5 m along the depth of the scene.

Here, β is the weighting factor for the importance of balance between position and orientation error. Although quaternions are preferred to be used as a representation for the orientation, there are cases where such a representation can be ambiguous too. Unit quaternions $\mathbf{q}$ and $-\mathbf{q}$ denote the same rotation, we cope with such a case as follows.

$$\phi_1(\mathbf{q_1}, \mathbf{q}_2) = \min\{\|\mathbf{q_1} - \mathbf{q_2}\|_2, \|\mathbf{q_1} + \mathbf{q_2}\|_2\} \tag{3}$$

The function in Eq. (3) is pseudometric than a metric on unit quaternions. Due to its nature of 2 to 1 mapping to $\mathrm{SO}(3)$, the pseudometric on the unit quaternions becomes a metric on 3D rotations [6]. It is important to regularize neural networks while training as it prevents over fitting. We use L_2 regularization which penalizes the values of the network parameters θ and help the network to generalize. We regularize only the weights of the network using Eq. (4) and not the biases. λ is a hyperparameter that controls the regularization strength.

$$\mathcal{L}(\theta) = \mathcal{L}(\theta) + \lambda \|\theta\|_2 \tag{4}$$

C. Synthetic Viewpoint Generation

Training deep neural networks require large amount of training data. We leverage the idea of transfer learning to train our network. As our task is to regress a global 6-DoF camera pose in a map, we propose to augment the pose space in 3D and generate the corresponding images and pose labels from the original RGB training images.

For image classification tasks, general data augmentation methods include color and shape distortions in 2D image space as it does not affect the class label. In our case random 2D shape augmentations cannot be applied as it would affect the pose of the camera. Therefore, for pose space coverage, we create synthetic viewpoints in 3D from

the training images and its associated synthetic depth. We do not have any stereo information, so we generate depth images from a single RGB image during the training phase. We use the method of Liu et al. [11] to generate depth images. We assume a pinhole camera model in our approach, that defines the relationship between a 3D point $\mathbf{p} = (x, y, z)^T \in \mathbb{R}^3$ and a 2D pixel position $x = (i, j)^T \in \mathbb{R}^2$.

$$\pi(x, y, z)^T = \left(\frac{f_x x}{z} + c_x, \frac{f_y y}{z} + c_y\right) = (i, j)^T \quad (5)$$

Here, f_x, f_y, c_x, c_y refer to the focal length and the optical center of the camera respectively. Given the depth z of a pixel (i, j), we can reconstruct the 3D point as follows:

$$\rho(i, j, z) = \left(\frac{(i - c_x)z}{f_x}, \frac{(j - c_y)z}{f_y}, z\right)^T \quad (6)$$

We generate a local 3D point cloud using Eq. (6). Then we apply 6 different pose variations. We rotate the camera around pitch, yaw by $\pm 5^\circ$ as strong rotations around roll are not expected. For more spatial coverage, we also synthesize views at ± 0.5 m along the depth of the scene. The resulting depth maps from the single image-based CNN predictions are not highly accurate, hence it limits large variations for pose synthesis and would degrade the synthesized image quality.

Given the original rotation $\mathbf{R}$ of a point $\mathbf{p}$ in global coordinates, the translation of the camera from origin of the global coordinate system in the camera-centred coordinates is given by $\mathbf{t} = -\mathbf{R}\mathbf{p}$. We calculate the synthesized rotation $\mathbf{R_s}$, translation $\mathbf{t_s}$ and position $\mathbf{p_s}$ after applying $\Delta\mathbf{R}$ rotation around a certain axis and $\Delta\mathbf{t}$ translation using Eq. (7)-(9).

$$\mathbf{R_s} = (\Delta\mathbf{R})\mathbf{R} \quad (7)$$

$$\mathbf{t_s} = (\mathbf{R_s}\mathbf{R}^\mathbf{T})\mathbf{t} + \Delta\mathbf{t} \quad (8)$$

$$\mathbf{p_s} = -\mathbf{R_s}^\mathbf{T}\mathbf{t_s} \quad (9)$$

This gives us the new pose labels for the synthetic viewpoints. To generate synthetic images, we project these 3D points to 2D using Eq. (5). The resulting data augmentations are shown in Fig. 3. We train our network with the original training images and the synthesized images. During the localization phase, our approach just uses a monocular RGB image and estimates the global 6-DoF camera pose with a constant time complexity of $\mathcal{O}(1)$.

IV. Experiments

We have evaluated our approach on four public benchmark datasets (Kings College, ShopFacade, St Mary Church, Old Hospital) to show the robustness of our approach in outdoor scenarios. These datasets comprise of images recorded from a hand-held camera in London. In this section, we will discuss the quantitative evaluation of different aspects of our approach on these datasets. We report positional errors in meters and the orientation errors in degrees. To train our network, we use a mini-batch of 64 images, learning rate of 10^{-2} and $\lambda = 0.1$ for the L_2 regularization of the weights. Furthermore, to show that our approach generalizes well to

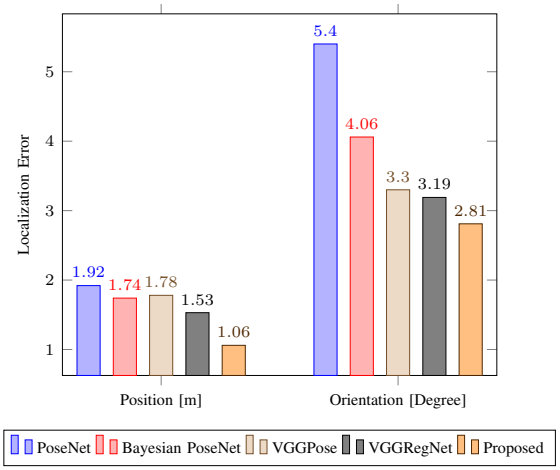

Fig. 4: Our proposed regression architecture and the data augmentation method in the 3D pose space outperforms vanilla VGG16, PoseNet and Bayesian PoseNet. These errors are reported for Kings College dataset.

different scenes, we use the same parameters to train our network. Both the methods (PoseNet, Bayesian PoseNet) use grid search to optimize the main tuning parameter β for each dataset. In our experiments, the relatively similar value of $\beta(200, 250)$ for all the scenarios shows the generalization of our method with a small variation on the accuracy. As extensive hyperparameter optimization of our approach will lead to better results, we regard the presented results as the lower bound for the accuracy. We denote the vanilla VGG16 architecture with pose regressors (fc9, fc10) as VGGPose, the proposed segregated architecture as VGGRegNet, and the proposed architecture trained with data augmentation as Proposed. We present the median errors for both orientation and position and also the average error over all the datasets. The ground truth poses are available from the 3D reconstructed models of the datasets that are created using SfM. We calculate L_2 norm of difference in positions for positional accuracy. The orientation errors are reported in degrees by calculating the difference between the estimated and ground truth quaternions as $2\arccos(|q_1 \cdot q_2|)\frac{180}{\pi}$.

We first discuss the advantage of using our proposed architecture and the data augmentation on the Kings College dataset. Fig. 4 shows the comparison of all the methods. VGGPose achieves 3.89% better accuracy in orientation and 7.29% better accuracy in position than PoseNet. This exhibits the potential of VGGPose as a better pose regressor. The reduced trainable parameters enable VGGPose to learn a better discriminative function for camera pose estimation. We also quantify the gain in accuracy of our proposed architecture. VGGRegNet results in further 14% reduction in position error and 3.3% reduction in the angular deviations.

Next, we compare our approach to PoseNet and Bayesian PoseNet over all 4 datasets. Fig. 7 shows median localization errors for both position and orientation on these datasets. Kings College has the largest spatial extent of 5000 m^2 amongst all the datasets. It consists of 1220 training images and 346 testing images. It does not contain extreme angular

Fig. 5: This figure shows challenging test images from all the datasets. Images are subjected to severe occlusions, camera rotations and structural ambiguities. From *Left:* Kings College, Old Hospital, Shop Facade, St. Mary Church.

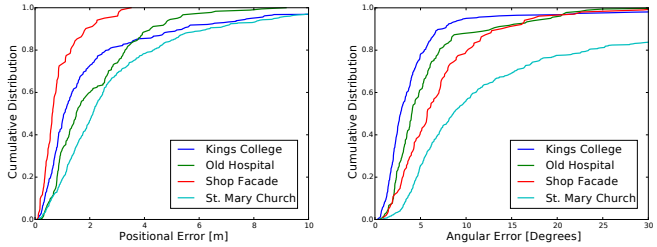

Fig. 6: Comparison of the localization accuracy on all the datasets as a cumulative histogram exhibits the relative errors between them. It can be observed that St. Mary Church is the most challenging dataset amongst them. Shop Facade has a relatively small spatial extent, hence its positional localization errors are least of them all.

deviations as the images are recorded while walking along the college. Our approach outperforms all state-of-the-art methods with positional and rotational errors are combined. ShopFacade dataset contains large angular variations and does not cover large space spatially as it covers an area of $875\,\mathrm{m}^2$. Our approach outperforms Bayesian PoseNet with the gain of 42.5% and 24% accuracy in position and orientation respectively. Fig. 3 shows that data augmentation helps to cover more orientations from a single training image in 3D space. This leads to better regression as the pose regressors have more data points for association and the weighted mean of these data points would give better pose estimate.

Old Hospital has a spatial extent of $2000\,\mathrm{m}^2$, it contains 895 training and 182 testing images. It shows relatively larger positional errors than Kings College. It is due to the large translational deviations of the camera along the depth of the scene. Our data augmentation along the zoom level and pitch and yaw makes our approach robust towards these deviations. We achieve 35% more accurate positional estimates and our orientation predictions are 27.5% more robust than Bayesian PoseNet. This dataset also contains strong angular deviations around pitch along with substantial changes in the position of the camera as well. In such scenarios, one could also exploit combined data augmentation, where the synthetic image not only undergoes a single variation in either position or rotation but is synthesized from the combination of both of these. We leave this investigation for future work.

The fourth dataset is recorded by walking a complete loop around St. Mary Church. It contains 1487 training images and 530 testing images. This dataset proved to be one of the most challenging dataset for pose regression. It comprises of strong camera rotations while covering a spatial area of $4800\,\mathrm{m}^2$. The Church has many similar windows on its

Error	PoseNet	Bayesian PoseNet	Proposed
Positional Error [m]	2.1	1.9	**1.3**
Angular Error [Deg]	6.8	6.3	**5.2**

TABLE I: Average localization errors over all the datasets.

periphery which makes the pose estimation ambiguous as shown in Fig. 5. Our approach perform as good as Bayesian PoseNet for positional accuracy and reduces the median orientation by 3.22%. Combined data augmentation might be helpful in such scenarios or probabilistic modeling of the resulting pose estimates would also help in such cases as indicated by the improvement of Bayesian PoseNet over PoseNet.

We also show the cumulative localization errors for both position and orientation over all the datasets in Fig. 6. It can be observed from the figure that positional localization errors for areas with larger spatial extent are relatively higher than those which cover relatively smaller areas. It is not only the spatial extent but also the magnitude of camera translations in the area. Although, Old Hospital has a smaller spatial extent, it has relatively lower positional localization accuracy than Kings College because of large spatial camera movements. The cumulative distributions of angular errors show that datasets with large angular deviations (Shop Facade and St. Mary Church) resulted in higher orientation errors than scenarios where the camera did not undergo severe rotations. We also report the average localization accuracy over all the datasets in Table I. Our position estimates are 31.6% and orientation estimates are 17.4% more accurate than Bayesian Posenet. Regarding timing, our approach only requires a single forward pass through the network for pose prediction, taking 12.5 ms on NVIDIA-TITAN X and it is independent of the database size.

Note that, a very recent method by Walch et al. [24] that uses spatial-LSTM combined with PoseNet architecture shows similar average positional error of 1.3 m and higher average rotational error of $5.5^\circ(5.45\%)$ than our proposed method. We believe that learning spatial context of image features using LSTMs is an interesting idea and would lead to further improvement in the localization accuracy. We do not report results on the *Street* dataset, which is the final outdoor dataset from Cambridge Landmarks. The same observation is pointed out by Walch et al. about this peculiar behavior on this dataset. This dataset comprises of videos recorded in opposite compass directions with similar spatial positions resulting in large angular deviations at similar global position. As we do not perform grid search for tuning parameters, the model could not converge with the same tuning parameters.

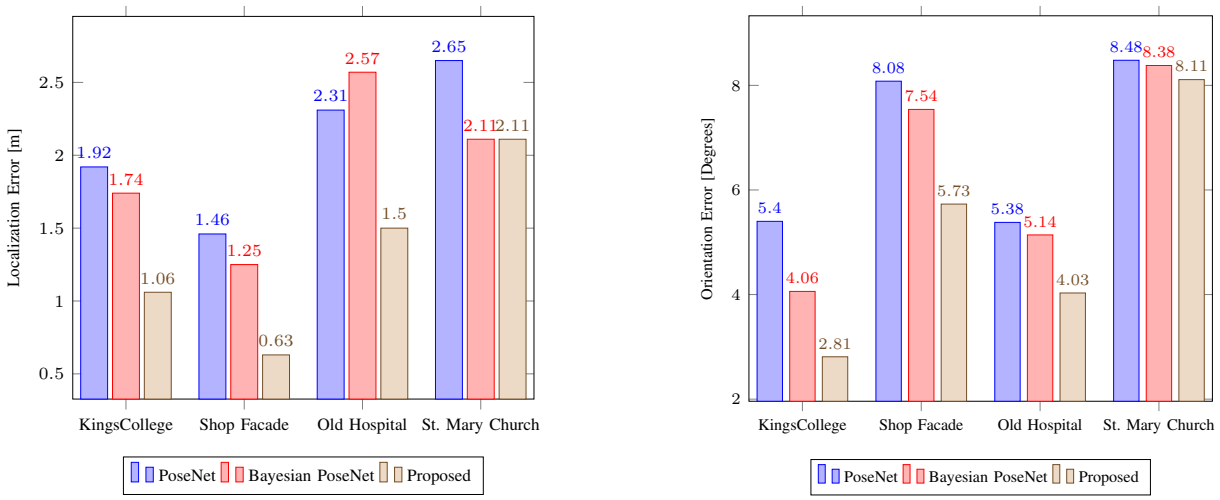

Fig. 7: Our approach outperforms PoseNet and Bayesian PoseNet on these datasets. using the same parameter settings for each of them. This exhibits that our approach generalizes well to different scenes.

V. Conclusion

In this paper, we proposed a novel approach for regressing 6-DoF camera poses with a mononcular RGB image in outdoor environments using convolutional neural networks. We proposed a novel architecture for deep regression and showed its effectiveness compared to state-of-the-art methods. The proposed data augmentation in 3D pose space resulted in a substantial improvement in the localization accuracy. In extensive experiments we demonstrated that our approach outperformes state-of-the-art CNN-based methods for metric localization in outdoor scenarios. For future work we plan to investigate the scalability of such networks for city scale pose regression and generalization across large time lags where methods based on point features do not perform well.

References

[1] H. Badino, D. Huber, and T. Kanade. Real-time topometric localization. In *Proc. of the IEEE Int. Conf. on Robotics and Automation (ICRA)*, 2012.

[2] Tim Caselitz, Bastian Steder, Michael Ruhnke, and Wolfram Burgard. Monocular camera localization in 3d lidar maps. In *Int. Conf. on Intelligent Robots and Systems (IROS)*, pages 1926–1931. IEEE, 2016.

[3] Jakob Engel, Thomas Schöps, and Daniel Cremers. Lsd-slam: Large-scale direct monocular slam. In *Proc. of the European Conf. on Computer Vision*, pages 834–849. Springer, 2014.

[4] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Aistats*, volume 9, pages 249–256, 2010.

[5] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[6] Du Q Huynh. Metrics for 3d rotations: Comparison and analysis. *Journal of Mathematical Imaging and Vision*, 35(2):155–164, 2009.

[7] Alex Kendall and Roberto Cipolla. Modelling uncertainty in deep learning for camera relocalization. In *Proc. of the IEEE Int. Conf. on Robotics and Automation (ICRA)*. IEEE, 2016.

[8] Alex Kendall, Matthew Grimes, and Roberto Cipolla. Posenet: A convolutional network for real-time 6-dof camera relocalization. In *Int. Conf. on Computer Vision (ICCV)*, pages 2938–2946, 2015.

[9] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[10] Yunpeng Li, Noah Snavely, Daniel P Huttenlocher, and Pascal Fua. Worldwide pose estimation using 3d point clouds. pages 147–163, 2016.

[11] Fayao Liu, Chunhua Shen, and Guosheng Lin. Deep convolutional neural fields for depth estimation from a single image. In *Proc. of the IEEE Int. Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2015. URL `http://arxiv.org/abs/1411.6387`.

[12] D.G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, pages 91–110, November 2004.

[13] Stephanie Lowry, Niko Sünderhauf, Paul Newman, John J Leonard, David Cox, Peter Corke, and Michael J Milford. Visual place recognition: A survey. *IEEE Transactions on Robotics*, 32(1):1–19, 2016.

[14] Colin McManus, Ben Upcroft, and Paul Newmann. Scene signatures: Localised and point-less features for localisation. In *Proc. of Robotics: Science and Systems (RSS)*, 2014.

[15] M. Milford and G. Wyeth. Seqslam: Visual route-based navigation for sunny summer days and stormy winter nights. In *Proc. of the IEEE Int. Conf. on Robotics and Automation (ICRA)*, 2012.

[16] Raul Mur-Artal, Jose Maria Martinez Montiel, and Juan D Tardos. Orb-slam: a versatile and accurate monocular slam system. *IEEE Transactions on Robotics*, 31(5):1147–1163, 2015.

[17] Peer Neubert and Peter Protzel. Beyond holistic descriptors, keypoints, and fixed patches: Multiscale superpixel grids for place recognition in changing environments. *IEEE Robotics and Automation Letters*, 1(1): 484–491, 2016.

[18] Geoffrey Pascoe, William Maddern, and Paul Newman. Direct visual localisation and calibration for road vehicles in changing city environments. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 9–16, 2015.

[19] Torsten Sattler, Bastian Leibe, and Leif Kobbelt. Fast image-based localization using direct 2d-to-3d matching. In *Int. Conf. on Computer Vision (ICCV)*, pages 667–674. IEEE, 2011.

[20] Torsten Sattler, Bastian Leibe, and Leif Kobbelt. Efficient & effective prioritized matching for large-scale image-based localization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2016.

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

[22] Niko Sünderhauf, Sareh Shirazi, Adam Jacobson, Edward Pepperell, Feras Dayoub, Ben Upcroft, and Michael Milford. Place recognition with convnet landmarks: Viewpoint-robust, condition-robust, training-free. In *Proc. of Robotics: Science and Systems (RSS)*, 2015.

[23] Akihiko Torii, Relja Arandjelovic, Josef Sivic, Masatoshi Okutomi, and Tomas Pajdla. 24/7 place recognition by view synthesis. In *Proc. of the IEEE Int. Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 1808–1817, 2015.

[24] Florian Walch, Caner Hazirbas, Laura Leal-Taixé, Torsten Sattler, Sebastian Hilsenbeck, and Daniel Cremers. Image-based localization with spatial lstms. *arXiv preprint arXiv:1611.07890*, 2016.

[25] Bernhard Zeisl, Torsten Sattler, and Marc Pollefeys. Camera pose voting for large-scale image-based localization. In *Proc. of the IEEE Int. Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[26] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. Learning deep features for scene recognition using places database. In *Advances in neural information processing systems*, 2014.