Deep learning based camera pose estimation in multi-view environment

Jorge L. Charco[1,3], Boris X. Vintimilla[1], Angel D. Sappa[1,2]

[1]Escuela Superior Politécnica del Litoral, ESPOL,
Facultad de Ingeniería en Electricidad y Computación, CIDIS,
Campus Gustavo Galindo Km. 30.5 Va Perimetral, P.O. Box 09-01-5863, Guayaquil, Ecuador

[2]Computer Vision Center, Edi ci O, Campus UAB,
08193, Bellaterra, Barcelona, Spain

[3] University of Guayaquil, Faculty of Mathematics and Physical Sciences,
P.B. EC090514, Guayaquil, Ecuador

{jlcharco, boris.vintimilla, asappa}@espol.edu.ec

***Abstract*—This paper proposes to use a deep learning network architecture for relative camera pose estimation on a multi-view environment. The proposed network is a variant architecture of AlexNet to use as regressor for prediction the relative translation and rotation as output. The proposed approach is trained from scratch on a large data set that takes as input a pair of images from the same scene. This new architecture is compared with a previous approach using standard metrics, obtaining better results on the relative camera pose.**

***Index Terms*—Deep learning, Camera pose estimation, Multi-view environment, Siamese architecture.**

I. Introduction

Camera calibration is a process that allows getting intrinsic (i.e., camera matrix, focal distance, distortion) and extrinsic parameters (rotation and translation) of a camera from a special calibration patterns. Different algorithms of computer vision do this procedure to have correspondences between 3D world points and 2D images points on camera plane. These correspondences are later on used to obtain the calibration parameters by means of an energy minimization process.

Computer vision has some difficulties when trying to solve this challenging camera calibration problem since many factors affect this process, such as illumination, low resolution and few image features. As mentioned above, image correspondences, which are based on feature detection, is the central point in any camera calibration process. Different approaches have been proposed in the literature for feature point detection and matching, for instance SURF [1], ORB [2], SIFT [3] just to mention a few. Unfortunately, these algorithms have low accuracy when there is not enough feature points to be matched.

Recently, convolutional neural networks (CNN) have been used to improve state-of-art results in tasks such as images classification and segmentation, pattern recognition and images enhancement. Nowadays, these methods are widely used due to their high precision in different spectrum such as infrared and visible. CNN based camera calibration algorithms have been also recently proposed. They can be classified into two categories: single and multi-view environment. The single-view approach is composed of monocular images that capture environment from the same angle and camera in a sequence of images, for which could be occluded an important feature depending on the camera angle, becoming an important problem to solve. In [4] the authors have proposed a robust CNN architecture, which can be used in real-time monocular six degree of freedom environment. The approach is used to obtain camera pose from a single RGB image, even with difficult lighting, motion blur and different camera intrinsic parameters. In [5] an updated version of the previous approach is proposed. The authors propose a similar architecture but with a new loss function to learn camera pose. On the other hand, the multi-view approaches solve the problem of occluded features since the scene is observed from different angles at the same time. Obviously, there must be a considerable overlap between the images. It should be noticed that all these learning based processes have a main limitation, which is related with the data set needed to train the network. Additionally, also related with the size of the data set, is the required computational cost. In the current paper a multi-view approach based on the usage of a CNN architecture is proposed to estimate relative extrinsic camera parameters between two images of the same scene. The training process is performed using the DTU Robot Image Dataset [6], which contains a large set of image pairs with their relative pose information. Different training strategies are evaluated to get the better results. The rest of the paper is organized as follows. In Section II previous works on relative camera pose estimation are summarized; this is followed by Section III, which presents the proposed approach to obtain extrinsic camera parameters. The results of conducted experiments are reported in Section IV together with a detailed description of the used dataset; and finally, Section V concludes this article.

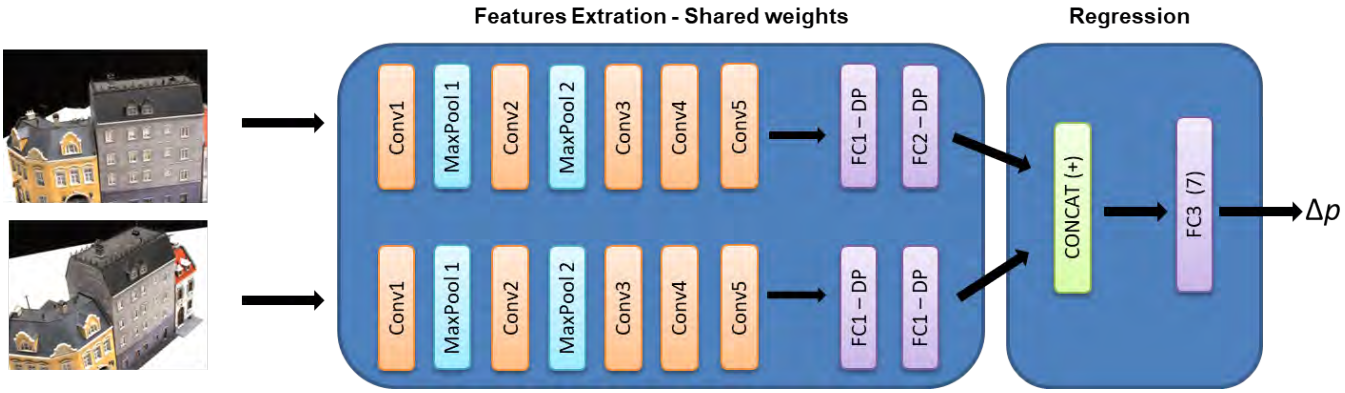

Fig. 1. Siamese architecture is feed with two input images that capture a scene from different angles. The regression part contains three fully-connected and dropout to estimate the extrinsic parameters of two cameras. A 7-dimensions vector is obtained as output.

II. Related Work

Over time, many solutions have been implemented to solve the problems of camera calibration. In [7], the authors propose a calibration process that does not require a calibration object with known 3D shape. It only requires point matches of image sequences. These points of interest are tracked through the images as the camera moves. The approach is based on two steps; the first one find the epipolar transformation while the second one uses Kruppa equations to estimate the camera parameters. A particular approach has been presented in [8]. The authors propose a method that is based on the usage of at least three images of a scene. These images are obtained from the same point in space but with different camera's orientation. In this approach there is not epipolar structure since all the images are taken from the same point in space, which make the point correspondences easier. On the contrary to the previous approaches, traditional methods are based on feature point detection and mathcing (e.g., SURF [1], ORB [2], SIFT [3]). As mentioned above, the main disadvantage of these methods is related with the number of corresponding 2D image points needed. This problem become harder with there is a lack of illumination or texture. In [9] a comparative evaluation of classical feature point descriptors in infrared and visible spectrum is presented. The robustness to changes in rotation, scaling, blur, and additive noise has been evaluated, obtained similar results between both spectrum. The SIFT algorithm has been used by [10] to calibrate the stereo cameras. For this, the first camera is calibrated using a plane based chessboard and after calibrated the intrinsic parameters of the two cameras are obtained. With this information, the essential matrix is obtained together with translation and rotation matrix. A method that uses the techniques described above has been presented in [11] for tackling the structure from motion (SfM) problem (i.e., reconstruction of 3D-images from a video sequence). In the SfM problem, camera position is recovered from the extracted feature points. In [12] the authors propose to solve the SfM by means of bundle adjustment algorithm, which uses curves partially observed in all images to refine camera position estimation.

In last years, many solutions using CNNs have been applied in computer vision problems considering the powerful to extract features on images. A few works have been proposed in the context of relative camera pose estimation. One of the work in this domain has been presented by [13]. The autors propose a relative pose estimation between two cameras using a Siamese CNN architecture. It was trained with image pairs from same scene. The architecture was trained using transfer learning from a large scale classification data set. An Euclidean loss function is used to estimate the relative translation and rotation as output. Each branch is composed of convolutional layer and activation function (ReLU). They have two fully-connected (FC1 and FC2) layers that estimate translation and rotation respectively.

III. Proposed Approach

Figure 1 shows the approach propose to estimate relative camera pose by using two images of same scene. Camera pose is represented by a 7-dimension vector: $\Delta p = [\hat{t}, \hat{r}]$, where $\hat{t}$ is a 3-dimensions vector that represents the translation and $\hat{r}$ is a 4-dimensions vector (quaternions) that represents the rotation. The proposed approach is a Siamese CNN architecture that contains two identical branches with shared weight. They are composed of convolutional layers, pooling and rectified linear unit (ReLU) as activation function. Additional, two fully-connected layers are proposed with hyperbolic tangent (tanh) as activation function. In detail, the proposed approach is based a modification of AlexNet architecture [14] ; more details are as follows:

- Replace all fully-connected and softmax classifier to output a 7-dimensions vector.
- Append two fully-connected layers before the final regressor of feature size 1024 each layer and dropout.
- Normalized to unit length the prediction at test time.

The input image of proposed approach was resized to the 240x320 pixels for training and testing. Image enhancement function (1) is used to improve the features of the images before of training phase:

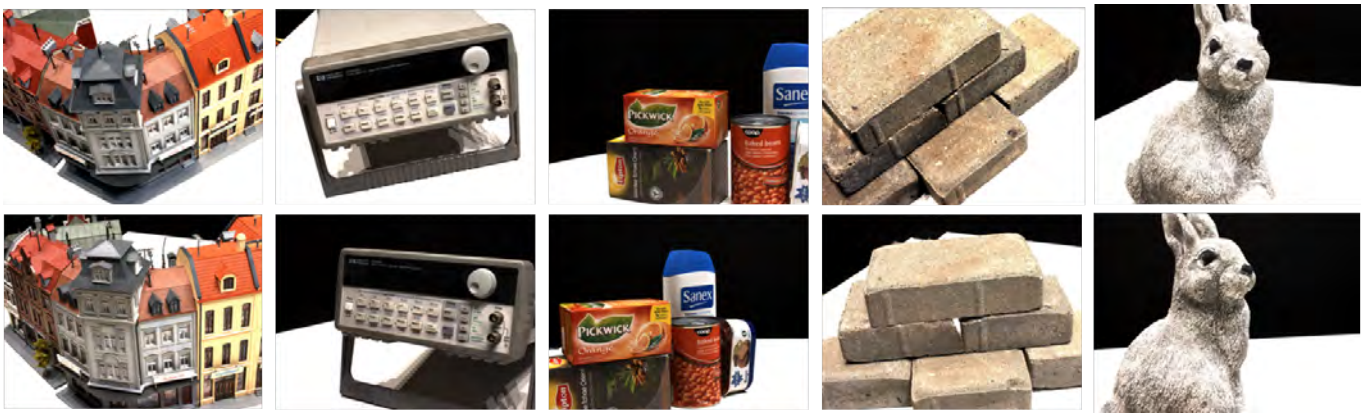

Fig. 2. DTU Robot Image Dataset [6]. Some of the image pairs used to train and evaluate the proposed approach (ground truth of camera pose is estimated through robotic arm).

$$\mathcal{I}_{(r,c)} = \left[\frac{\mathcal{I}_{(r,c)} - \mathcal{I}_{(r,c)MIN}}{\mathcal{I}_{(r,c)MAX} - \mathcal{I}_{(r,c)MIN}}\right][MAX - MIN] + MIN \quad (1)$$

where $\mathcal{I}_{(r,c)}$ corresponds to an image's pixel, $\mathcal{I}_{(r,c)MIN}$ and $\mathcal{I}_{(r,c)MAX}$ are the minimum and maximum values in the image respectively. While MAX and MIN correspond to 0 and 255 respectively.

The fully-connected layers allow learning non-linear combinations of high-level features of convolutional layers. Dropout is applied to avoid overfitting. The proposed approach allows to train two image pairs. The output of both branches are concatenated to estimate the prediction of 7-dimensions vector that is represented as translation (3-dimensions) and rotation (4-dimensions) (See Fig. 1). The proposed approach was trained from scratch with an Euclidean loss function.

$$\mathcal{L}_{final}(I) = \|t - \hat{t}\|_\gamma + \|r - \hat{r}\|_\gamma \quad (2)$$

where γ is $L2$ euclidean norm. We normalize the ground truth relative translation and rotation to unit length. Therefore $\|r\| = \|t\| = 1$. Hence a normalization stage is added at training and testing phase

IV. Experiments Results

The proposed approach has been evaluated using image pairs of the same scene and their corresponding extrinsic parameters obtained from [6]. Tensorflow library [15] and python [16] has been used for implementation of the proposed approach. It has been trained using a 2.5 Ghz. dual core with 96GB of memory and Tesla K20m GPU. The Adam algorithm [17] has been used as optimizer with a learning rate of 10^{-4}, reduced by 5% every epoch during 40 epochs. Every training process took approximately about 49 hours using a batch of 100.

A. Dataset

In order to train the proposed architecture a large data set is necessary. In the current work pairs of images from from [6], referred to as DTU, has been used. In this data set ground truth of camera pose (translation and rotation) has been obtained by using a robotic arm. The DTU data set consists of 128 scenes covering 64 different camera positions. The original images have fixed-size of 1200x1600 pixels. In order to evaluate the proposed model, the data set has been split up into training, validation and testing sets. Only the images pairs that have overlapping in the field view are considered. The training set is composed of 59800 image pairs. Additional, the validation and testing sets are composed of 11200 and 3700 image pairs respectively. In Fig. 2) five pairs of images are presented.

As preprocessing step, DTU data set has been resized from 1200x1600 pixels to 240x320 pixels, keeping aspect ratio of original images. Image enhancement function (1) has been used before the training step. This helped to improve the features of the images (see Fig 3). The ground truth of each image pairs contains a 7-dimensions vector, i.e., translation (3-dimensions) and rotation represented by quaternion (4-dimensions). Additionally, each image pairs and their ground truth camera pose (translation and rotation) has been saved into TFRecord file of tensorflow to better manage memory during training, validation and testing steps.

B. Results and comparisons

Different quantitative metrics has been used to evaluate the proposed approach. One of the metrics is the angular error (AE) computed between the obtained result and ground truth value. AE is computed only on the four elements of the quaternion vector [18]. Additionally, The euclidean distance (ED) is also considered a quantitative evaluation to determine the distance between translations vectors. For a fair comparison, the cnnB architecture [13] has been trained with the same parameters and data set without transfer learning. Figure 4 and Fig. 5 show comparisons of the results obtained with cnnB and the proposed approach.

The presented results confirm that the proposed approach obtains better accuracy than the cnnB architecture [13]. The improvements on the proposed approach correspond to the fully-connected layers that have been included to the feature

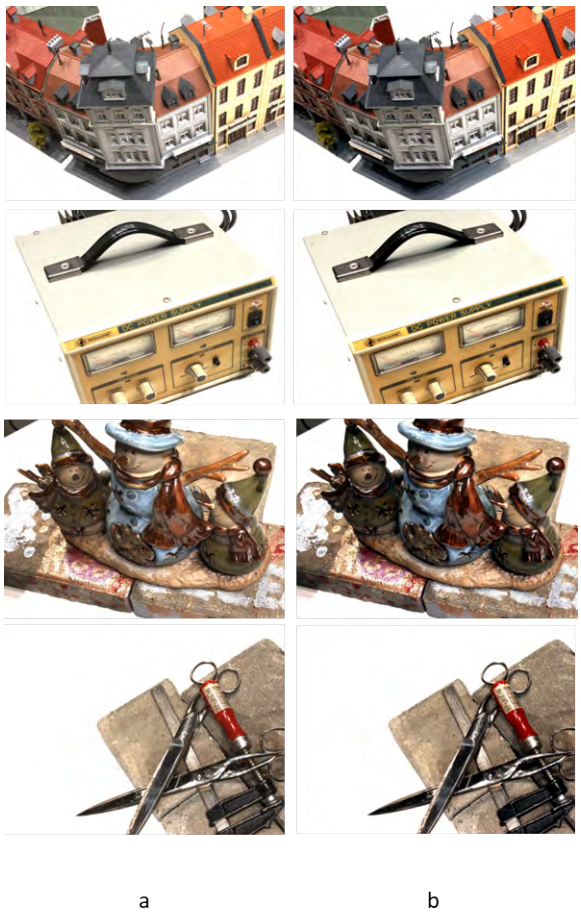

Fig. 3. (a) Result obtained after resizing original images to 240x320 pixels. (b) Enhanced images after applying the enhanced function (note how image are more colourful after this enhancement stage).

TABLE I
COMPARISON BETWEEN THE PREDICTION OBTAINED WITH OUR MODEL AND CNNB MODEL USING SIX DIFFERENT SCENES FROM THE TESTING SET.

Scene #	Our model		cnnB	
	Translation (ED)	Rotation (AE)	Translation (ED)	Rotation (AE)
84	0.019	22.71°	0.141	88.77°
85	0.020	23.55°	0.135	80.51°
98	0.021	21.79°	0.132	77.41°
106	0.023	21.97°	0.155	75.04°
110	0.023	26.40°	0.111	76.39°
128	0.033	27.52°	0.143	92.90°

extraction stage; additionally, the combined use of branch output in the regressor stage helps to obtain better results (see Fig 1).

A set of six scenes, randomly selected, is chosen to evaluate with more details the proposed approaches (see Fig 6). Each scene contains 64 images taken with different camera positions. From this images about 70 pairs are obtained for the evaluation. Table I shows the average results obtained in each

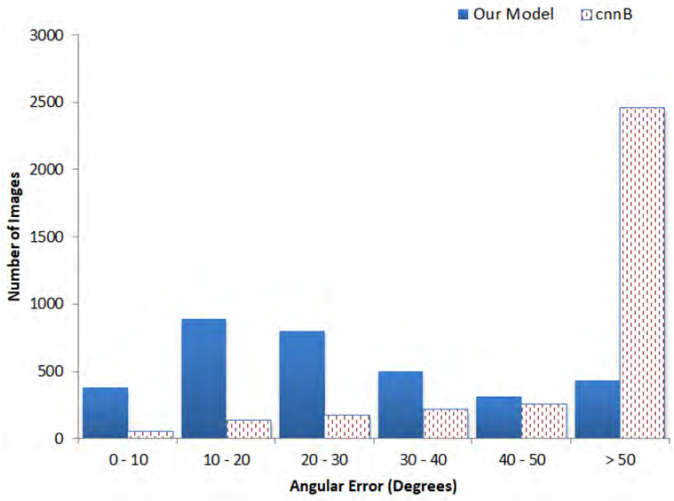

Fig. 4. Comparison of rotation error between proposed approach and approach proposed by [13] on DTU Image Robot Dataset.

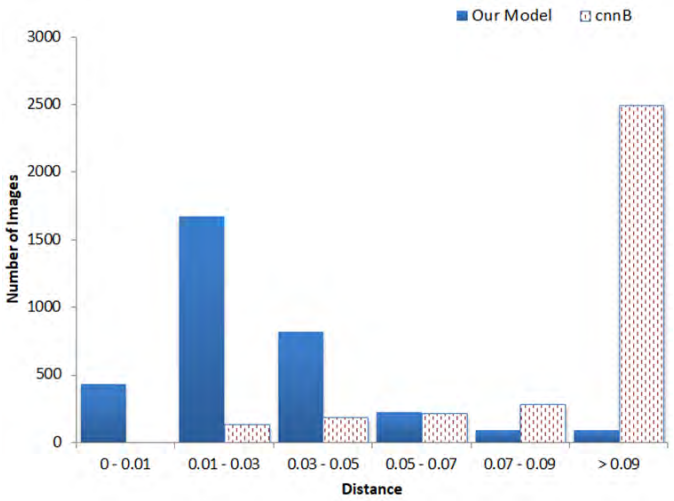

Fig. 5. Comparison of translation error between proposed approach and approach proposed by [13] on DTU Image Robot Dataset.

scene. According to these results, in all the scenes the proposed approach has smaller angular and distance errors than the cnnB architecture. The results with large translation and rotation values correspond a scene with lack of illumination or poor texture (e.g., scene 100, 128). On the contrary, the results improve significantly when the images have good conditions of lighting and features (e.g., scene 84, 98).

V. CONCLUSION

This paper addresses the challenging problem of estimating relative camera pose from two different images of the same scene by using a Siamese convolutional neural network. Experimental results show that the proposed approach performs a good camera pose estimation in most of the scenes. Accuracy is affected by images with lack of texture or poor illumination.

Fig. 6. DTU Robot Image Dataset [6]. Set of six random scenes from the testing set. The images were taken in different camera positions.

Future work will be focused on evaluating different models based on Siamese architecture; additionally, other loss functions will be implement. Finally, increasing the size of data set, including indoor or outdoor environments, will be considered.

ACKNOWLEDGMENT

This work has been partially supported by: the ESPOL project PRAIM (FIEC-09-2015); the Spanish Government under Projects TIN2014-56919-C3-2-R and TIN2017-89723-P; and the CERCA Programme / Generalitat de Catalunya". The authors gratefully acknowledge the support of the CYTED Network: Ibero-American Thematic Network on ICT Applications for Smart Cities (REF-518RT0559). The first author has been supported by Ecuador government under SENESCYT scholarship contract CZ05-000040-2018.

REFERENCES

[1] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (surf)," *Computer vision and image understanding*, vol. 110, no. 3, pp. 346–359, 2008.

[2] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "Orb: An efficient alternative to sift or surf," in *Computer Vision (ICCV), 2011 IEEE international conference on*. IEEE, 2011, pp. 2564–2571.

[3] E. N. Mortensen, H. Deng, and L. Shapiro, "A sift descriptor with global context," in *Computer vision and pattern recognition, 2005. CVPR 2005. IEEE computer society conference on*, vol. 1. IEEE, 2005, pp. 184–190.

[4] A. Kendall, M. Grimes, and R. Cipolla, "Posenet: A convolutional network for real-time 6-dof camera relocalization," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2938–2946.

[5] A. Kendall, R. Cipolla *et al.*, "Geometric loss functions for camera pose regression with deep learning," in *Proc. CVPR*, vol. 3, 2017, p. 8.

[6] R. Jensen, A. Dahl, G. Vogiatzis, E. Tola, and H. Aanæs, "Large scale multi-view stereopsis evaluation," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2014, pp. 406–413.

[7] O. D. Faugeras, Q. T. Luong, and S. J. Maybank, "Camera self-calibration: Theory and experiments," in *Computer Vision — ECCV'92*, G. Sandini, Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 1992, pp. 321–334.

[8] R. I. Hartley, "Self-calibration from multiple views with a rotating camera," in *European Conference on Computer Vision*. Springer, 1994, pp. 471–478.

[9] P. Ricaurte, C. Chiln, C. A. Aguilera-Carrasco, B. X. Vintimilla, and A. D. Sappa, "Performance evaluation of feature point descriptors in the infrared domain," in *2014 International Conference on Computer Vision Theory and Applications (VISAPP)*, vol. 1, Jan 2014, pp. 545–550.

[10] R. Liu, H. Zhang, M. Liu, X. Xia, and T. Hu, "Stereo cameras self-calibration based on sift," in *2009 International Conference on Measuring Technology and Mechatronics Automation*, vol. 1, April 2009, pp. 352–355.

[11] J. L. Schonberger and J.-M. Frahm, "Structure-from-motion revisited," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4104–4113.

[12] I. Nurutdinova and A. Fitzgibbon, "Towards pointless structure from motion: 3d reconstruction and camera parameters from general 3d curves," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 2363–2371.

[13] I. Melekhov, J. Ylioinas, J. Kannala, and E. Rahtu, "Relative camera pose estimation using convolutional neural networks," in *International Conference on Advanced Concepts for Intelligent Vision Systems*. Springer, 2017, pp. 675–687.

[14] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2012, pp. 1097–1105. [Online]. Available: http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf

[15] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard *et al.*, "Tensorflow: a system for large-scale machine learning." in *OSDI*, vol. 16, 2016, pp. 265–283.

[16] G. Van Rossum *et al.*, "Python programming language." in *USENIX Annual Technical Conference*, vol. 41, 2007, p. 36.

[17] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[18] D. Q. Huynh, "Metrics for 3d rotations: Comparison and analysis," *Journal of Mathematical Imaging and Vision*, vol. 35, no. 2, pp. 155–164, 2009.