Ruhr-Universität Bochum

Bachelorarbeit

Pose Estimation in Gebäuden anhand von Convolutional Neural Networks und simulierten 3D-Daten

Schriftliche Prüfungsarbeit
für die Bachelor-Prüfung des Studiengangs Angewandte Informatik an der Ruhr-Universität Bochum

vorgelegt von
Abdullah Sahin

am
Lehrstuhl für Informatik im Bauwesen
Prof. Dr.-Ing Markus König

Abgabedatum:	April 4th, 2019
Matrikelnummer:	108016202304
1. Prüfer:	Prof. Dr.-Ing. Markus König
2. Prüfer:	Patrick Herbers, M. Sc.

Abstract

.....

Erklärung

Ich erkläre, dass das Thema dieser Arbeit nicht identisch ist mit dem Thema einer von mir bereits für eine andere Prüfung eingereichte Arbeit.

Ich erkläre weiterhin, dass ich die Arbeit nicht bereits an einer anderen Hochschule zur Erlangung eines akademischen Grades eingereicht habe.

Ich versichere, dass ich die Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen benutzt habe. Die Stellen der Arbeit, die anderen Werken dem Wortlaut oder dem Sinn nach entnommen sind, habe ich unter Angabe der Quellen der Entlehnung kenntlich gemacht. Dies gilt sinngemäß auch für gelieferte Zeichnungen, Skizzen, bildliche Darstellungen und dergleichen.

Statement

I hereby declare that except where the specific reference is make to the work of others, the contents of this thesis are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university.

This thesis is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text.

Datum Unterschrift

Inhaltsverzeichnis

1 Einleitung

Fügen Sie hier Ihren Text ein. Klicken Sie nach der Prüfung auf die farbig unterlegten Textstellen. oder nutzen Sie diesen Text als Beispiel für ein Paar Fehler , die Language-Tool erkennen kann: Ihm wurde Angst und bange. Mögliche stilistische Probleme werden blau hervorgehoben: Das ist besser wie vor drei Jahren. Eine Rechtschreibprüfun findet findet übrigens auch statt. Donnerstag, den 23.01.2019 war gutes Wetter. Die Beispiel endet hier.

2 Stand der Forschung und Grundlagen

Dieses Kapitel versucht einen Überblick des Forschungsstandes in den unterschiedlichen Aspekten der Arbeit zu verschaffen. Anschließend vermittelt das Kapitel notwendige Grundkenntnisse.

2.1 Einführung

Pose Estimation wird in dieser Arbeit als eine Methode der visuellen Lokalisierung *(Visual-Based Localization, kurz VBL)* betrachtet. VBL beschäftigt sich mit der Bestimmung der Pose *(Position + Orientierung)* eines visuellen Abfragematerials (z.B. ein RGB-Bild) in einer zuvor bekannten Szene [35]. Ein naheliegendes Themengebiet der Robotik ist die visuelle Ortswiedererkennung *(Visual Place Recognition, kurz VPR)* [25]. Die visuelle Ortswiedererkennung fokussiert sich auf das Feststellen eines bereits besuchten Ortes und definiert sich aus einer Datenverarbeitungs-, Kartografierungs- und einem Orientierungsmodul. Allgemein lässt sich das Prozess eines VPRs [25] folgend beschreiben. Die Daten werden vom Datenverarbeitungsmodul vorbereitet und anschließend an das Orientierungsmodul übergeben. Das Orientierungsmodul bestimmt die Position und/oder die Orientierung und entscheidet, auf dem Basis der immer aktuell gehaltenen Karte, ob ein Ort bereits besucht wurde. Eine interne Karte besuchter Orte wird durch das Kartografiemodul erstellt. Im Vergleich zur VPR [25] versucht die visuellen Lokalisierung [35] eine Pose zu bestimmen und benötigt daher kein Kartographierungsmodul.

Die rein visuellen Methoden des VBLs unterteilen sich in indirekte und direkte Methoden [25]. Die indirekten Methoden behandeln das Lokalisierungsproblem als eine Bildersuche in einer Datenbank, ähnlich wie das *Content Based Image Retrieval* [23] Problem. Diese Art von Methoden benötigen eine sehr große Datenbank und sind bei der Suche schwach gegenüber Abweichungen zu den Vergleichsbildern [25]. Es gibt drei Arten der direkten Methoden: (1) Abgleichen von Features zu Punktwolken (2) Pose Regression mit Tiefenbilder *(RGB-D Bilder)*. (3) Pose Regression nur mit Bildern.

Diese Ansätze haben Ähnlichkeiten zur den indirekten Methoden und benötigen eine repräsentative 3D-Punktwolke der Szene [35].

Diese Forschungsprojekte liefern mit 3D-Kameras gewünschte Resultate, jedoch sind diese Kameras nicht die Gängigsten im Alltag.

Convolutional Neural Networks *(CNN)* werden erfolgreich im Bereich des Maschinelles Sehens, wie z. B. bei der Klassifizierung von Bildern [22, 39, 13] sowie bei der Objekterkennung [10, 37, 9] eingesetzt. Ein verbreiteter Ansatz beim Entwurf von CNNs ist das häufig zweckentfremdende Feintunen *(fine-tune)* der Netzwerkarchitekturen, die z. B. für die Bildklassifizierung angesichts der Aufgaben von ImageNet [38] konstruiert wurden. Dieser Ansatz konnte beispielsweise erfolgreich in der Objekterkennung [9], Objektsegmentierung [21, 27], semantische Segmentierung [32, 12] und Tiefenbestimmung [24] verfolgt werden. Seit Kurzem werden CNNs auch in den Anwendungsgebieten der

Lokalisierung verwendet. Zum Beispiel verwenden Parisotto *et al.* [33] CNNs in Bezug auf das SLAM Problem. Melekhov *et al.* [29] schätzen anhand CNNs die relative Pose zweier Kameras. Constante *et al.* [5] und Wang *et al.* [46] setzen es im Bereich der visuellen Odometrie ein.

Geleitet von den *state-of-the-art* Lokalisierungsergebnissen der CNNs stellen Kendall *et al.* [20] den ersten Ansatz zu direkten Posebestimmung nur mit RGB-Bildern vor. PoseNet [20] ist die Modifikation der GoogLeNet [41] Architektur und zweckentfremdet es von der Bildklassifizierung zu einem Pose-Regressor. Trainiert mit einem Datensatz, bestehend aus Paaren von Farbbild und Pose, kann es die sechs Freiheitsgrade der Kameraposen in unbekannten Szenen mittels eines Bildes bestimmen. Dieser Ansatz benötigt weder Tiefenbilder der Szene noch eine durchsuchbare Bildgalerie. Im Vergleich zu den metrischen Ansätzen wie SLAM oder visuelle Odometrie liefert es eine weniger akkurate Pose. Es bietet jedoch eine hohe Toleranz gegenüber Skalierungs- und Erscheinungsänderungen des Anfragebildes an [35].

Es gibt mehrere Ansätze, die die Genauigkeit von PoseNet [20] übertreffen. Einen Fortschritt erhalten die Autoren von PoseNet [20] durch die hier [19] vorgestellte Anpassung ihres Models an einem Bayessian Neural Network [6, 26]. Dieselben Autoren erweitern PoseNet [20] mit einer neuen Kostenfunktion unter Berücksichtigung von geometrischen Eigenschaften [18]. Wlach *et al.* [45] und Clark *et al.* [4] setzen Long-Short-Term-Memory *(LSTM)* [14] Einheiten ein, um Wissen aus der Korrelation von Bildsequenzen zu gewinnen. Wu *et al.* [47] und Naseer *et al.* [31] augmentieren den Trainingsdatensatz. Wu *et al.* [47] stocken den vorhandenen Datensatz auf, indem sie die Bilder künstlich rotieren. Naseer *et al.* [31] erweitern zuerst über ein weiteres CNN den Datensatz um Tiefenbildern. Anschließend simulieren sie RGB-Bilder aus verschiedenen Viewpoints. Im Vergleich zu PoseNet [20] verwenden Müller *et al.* [30] und Melekhov *et al.* [28] eine andere Architektur. Das Modell von Müller *et al.* [30] basiert auf die SqueezeNet [16] Architektur. Melekhov *et al.* stellen HourglassNet [28], basierend auf einem symmetrischen Encoder-Decoder Architektur, vor. Brahmbhatt *et al.* [3] und Valada *et al.* [42, 43] binden zusätzliche Informationen wie z.B. visuelle Odometrie, GPS oder IMU ein.

Jedes dieser Ansätze benötigen annotierte Traininsdaten. Für die Erstellung solcher Daten wurden beispielsweise mit entsprechender Hardware ausgerüstete Trolleys [15], 3D-Kameras [17] oder SfM-Methoden [20] eingesetzt.

Simulierte 3D-Daten werden in der Literatur oft eingesetzt, um das manuelle Erzeugen und Annotieren von Daten umzugehen. Pishchulin *et al.* [36], Peng *et al.* [34], Su *et al.* [40] und Varol *et al.* [44] erzeugen ihren Trainingsdaten, indem sie virtuelle Objekte auf reale Hintergrundbildern platzieren. Pishchulin *et al.* [36] generieren Daten zwecks Personenerkennung und Bestimmung derer körperlicher Pose. Zuvor werden auf den vorhandenen Bildern die körperliche Pose der Personen bestimmt und daran deren 3D Modelle rekonstruiert. Anschließend werden die 3D-Modelle in ihrer Pose variiert auf reale Hintergrundbildern platziert. Peng *et al.* [34] erstellen Daten um Objekte auf realen Bildern zu detektieren. Von jeder Objektklasse werden 3D-Modelle auf einem Hin-

tergrundbild aus einer Sammlung gelegt. Su *et al.* [40] generieren einen großen Datensatz mit 3D-Modellen, um den Viewpoint von Objekten auf realen Bildern zu bestimmen. Bei dieser Datengenerierung wird jedes virtuelle Objekt auf zufällige Hintergrundbildern positioniert und mit unterschiedlichen Konfigurationen *(wie z.B. Beleuchtung)* gerendert. Varol *et al.* [44] erstellen künstliche Personen auf Bildern, um beispielsweise den menschlichen Körper in seine Glieder zu segmentieren. Dabei rendern sie zufällige virtuelle Personen mit zufälliger Pose auf beliebige Hintergrundbildern. Fanello *et al.* [8] rendert künstliche Infrarotbilder von Händen sowie Gesichtern zwecks Tiefenerkennung und Segmentierung der Hand in den einzelnen Fingern sowie des Gesichtes in Bereiche aus einem RGB-Bild. Dosovitskiy *et al.* [7] erlernen mit synthetischen Daten den optischen Fluss von Bildsequenzen. Hierbei werden auf Hintergrundbildern aus einer Sammlung mehrmals bewegte virtuelle Stühle platziert.

Motiviert von der Datengenerierung über 3D-simulierten Daten stellt Ha *et al.* [11] einen Ansatz zur Bild-basierte Lokalisierung in Gebäuden vor. Dieser Forschungsansatz generiert synthetische Daten aus einem Building Information Modeling *(BIM)*. Bei den Daten werden die durch das vortrainierte VGG Netzwerk [39] extrahierte Features als wesentlich erachtet und in einer Datenbank gepflegt. Ein reales Aufnahmebild im Gebäude lässt sich durch den Vergleich der Features lokalisieren. Acharya *et al.* [1, 2] erzeugen ebenso Trainingsdaten aus einem BIM, jedoch verwenden sie zur Lokalisierung keine Datenbank bedürftiges Verfahren, sondern bestimmen die Pose direkt über PoseNet [20]. Die Daten werden entlang eines Flugbahnes aus der Simulation eines Korridors gesammelt. Hierbei werden sich in der Realitätstreue vom karikaturistisch zu fotorealistisch hin über zu fotorealistisch-texturiert unterscheidende Daten erzeugt. Die besten Ergebnisse konnten die Autoren trainiert mit den Gradienten- und Kantenbilder der karikaturistischen Daten, getestet auf die Gradientenbilder der realen Aufnahmen, erzielen.

Im weiteren Verlauf des Kapitels werden einige grundlegende Themen erläutert. Zuerst wird die Lineare Faltung erklärt und die Verarbeitung eines Bildes über den Sobelfilter zum Gradientenbild ausgeführt. Danach wird ein vertieftes Wissen an CNN vermittelt und anschließend bekannte CNN Modelle näher erläutert.

2.2 Lineare Faltung

2.2.1 Sobelfilter

2.2.2 Gradientenbild

Bei der Erzeugung von Gradienten- bzw. Kantenbilder gehen einerseits wichtige Informationen im Hinblick auf das Ursprungsbild verloren andererseits bleiben wichtige Informationen wie z. B. die geometrische Struktur erhalten.

2.3 Feedforward Neural Network

2.3.1 Aktivierungsfunktionen

2.4 Convulutional Neural Network

2.4.1 Convolution Layer

2.4.2 Pooling Layer

2.4.3 Fully Connected Layer

2.5 Bekannte CNN Modelle

2.5.1 GoogLeNet

2.5.2 PoseNet

3 Methodology

4 Discussion

5 Conclusion

Literatur

[1] D. Acharya, K. Khoshelham, and S. Winter. BIM-PoseNet: Indoor camera localisation using a 3d indoor model and deep learning from synthetic images. 150:245–258, . doi: 10.1016/j.isprsjprs.2019.02.020.

[2] D. Acharya, S. Roy, K. Khoshelham, and S. Winter. MODELLING UNCERTAINTY OF SINGLE IMAGE INDOOR LOCALISATION USING a 3d MODEL AND DEEP LEARNING. .

[3] S. Brahmbhatt, J. Gu, K. Kim, J. Hays, and J. Kautz. Geometry-aware learning of maps for camera localization. pages 2616–2625. URL `http://openaccess.thecvf.com/content_cvpr_2018/html/Brahmbhatt_Geometry-Aware_Learning_of_CVPR_2018_paper.html`.

[4] R. Clark, S. Wang, A. Markham, N. Trigoni, and H. Wen. VidLoc: A deep spatio-temporal model for 6-DoF video-clip relocalization. URL `http://arxiv.org/abs/1702.06521`.

[5] G. Costante, M. Mancini, P. Valigi, and T. A. Ciarfuglia. Exploring representation learning with CNNs for frame-to-frame ego-motion estimation. 1(1):18–25. ISSN 2377-3766, 2377-3774. doi: 10.1109/LRA.2015.2505717. URL `http://ieeexplore.ieee.org/document/7347378/`.

[6] J. S. Denker and Y. LeCun. Transforming neural-net output levels to probability distributions. In R. P. Lippmann, J. E. Moody, and D. S. Touretzky, editors, *Advances in Neural Information Processing Systems 3*, pages 853–859. Morgan-Kaufmann. URL `http://papers.nips.cc/paper/419-transforming-neural-net-output-levels-to-probability-distributions.pdf`.

[7] A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. v. d. Smagt, D. Cremers, and T. Brox. FlowNet: Learning optical flow with convolutional networks. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2758–2766. IEEE. ISBN 978-1-4673-8391-2. doi: 10.1109/ICCV.2015.316. URL `https://ieeexplore.ieee.org/document/7410673/`.

[8] S. R. Fanello, T. Paek, C. Keskin, S. Izadi, P. Kohli, D. Kim, D. Sweeney, A. Criminisi, J. Shotton, and S. B. Kang. Learning to be a depth camera for close-range human capture and interaction. 33(4):1–11. ISSN 07300301. doi: 10.1145/2601097.2601223. URL `http://dl.acm.org/citation.cfm?doid=2601097.2601223`.

[9] R. Girshick. Fast r-CNN. URL `http://arxiv.org/abs/1504.08083`.

[10] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. URL `http://arxiv.org/abs/1311.2524`.

[11] I. Ha, H. Kim, S. Park, and H. Kim. Image-based indoor localization using BIM and features of CNN. doi: 10.22260/ISARC2018/0107. URL http://www.iaarc.org/publications/2018_proceedings_of_the_35th_isarc/image_based_indoor_localization_using_bim_and_features_of_cnn.html.

[12] C. Hazirbas, L. Ma, C. Domokos, and D. Cremers. FuseNet: Incorporating depth into semantic segmentation via fusion-based CNN architecture. In S.-H. Lai, V. Lepetit, K. Nishino, and Y. Sato, editors, *Computer Vision – ACCV 2016*, volume 10111, pages 213–228. Springer International Publishing. ISBN 978-3-319-54180-8 978-3-319-54181-5. doi: 10.1007/978-3-319-54181-5_14. URL http://link.springer.com/10.1007/978-3-319-54181-5_14.

[13] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. URL http://arxiv.org/abs/1512.03385.

[14] S. Hochreiter and J. Schmidhuber. Long short-term memory. 9(8):1735–1780. ISSN 0899-7667, 1530-888X. doi: 10.1162/neco.1997.9.8.1735. URL http://www.mitpressjournals.org/doi/10.1162/neco.1997.9.8.1735.

[15] R. Huitl, G. Schroth, S. Hilsenbeck, F. Schweiger, and E. Steinbach. TUMindoor: An extensive image and point cloud dataset for visual indoor localization and mapping. In *2012 19th IEEE International Conference on Image Processing*, pages 1773–1776. IEEE. ISBN 978-1-4673-2533-2 978-1-4673-2534-9 978-1-4673-2532-5. doi: 10.1109/ICIP.2012.6467224. URL http://ieeexplore.ieee.org/document/6467224/.

[16] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5mb model size. URL http://arxiv.org/abs/1602.07360.

[17] S. Izadi, D. Kim, O. Hilliges, D. Molyneaux, R. Newcombe, P. Kohli, J. Shotton, S. Hodges, D. Freeman, A. Davison, and A. Fitzgibbon. *KinectFusion: Real-time 3D Reconstruction and Interaction Using a Moving Depth Camera*. ISBN 978-1-4503-0716-1. URL https://www.microsoft.com/en-us/research/publication/kinectfusion-real-time-3d-reconstruction-and-interaction-using-a-moving-depth-camera/.

[18] A. Kendall and R. Cipolla. Geometric loss functions for camera pose regression with deep learning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6555–6564. IEEE, . ISBN 978-1-5386-0457-1. doi: 10.1109/CVPR.2017.694. URL http://ieeexplore.ieee.org/document/8100177/.

[19] A. Kendall and R. Cipolla. Modelling uncertainty in deep learning for camera relocalization. . URL http://arxiv.org/abs/1509.05909.

[20] A. Kendall, M. Grimes, and R. Cipolla. PoseNet: A convolutional network for real-time 6-DOF camera relocalization. pages 2938–2946. URL https://www.cv-foundation.org/openaccess/content_iccv_2015/html/Kendall_PoseNet_A_Convolutional_ICCV_2015_paper.html.

[21] I. Kokkinos. Pushing the boundaries of boundary detection using deep learning. URL http://arxiv.org/abs/1511.07386.

[22] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc. URL http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf.

[23] M. S. Lew, N. Sebe, C. Djeraba, and R. Jain. Content-based multimedia information retrieval: State of the art and challenges. 2(1):1–19. ISSN 15516857. doi: 10.1145/1126004.1126005. URL http://portal.acm.org/citation.cfm?doid=1126004.1126005.

[24] 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. pages 1119–1127. URL http://openaccess.thecvf.com/content_cvpr_2015/html/Li_Depth_and_Surface_2015_CVPR_paper.html.

[25] S. Lowry, N. Sünderhauf, P. Newman, J. J. Leonard, D. Cox, P. Corke, and M. J. Milford. Visual place recognition: A survey. 32(1):1–19. ISSN 1552-3098. doi: 10.1109/TRO.2015.2496823.

[26] D. J. C. MacKay. A practical bayesian framework for backprop networks. 4:448–472.

[27] K.-K. Maninis, J. Pont-Tuset, P. Arbeláez, and L. Van Gool. Convolutional oriented boundaries. 9905:580–596. doi: 10.1007/978-3-319-46448-0_35. URL http://arxiv.org/abs/1608.02755.

[28] I. Melekhov, J. Ylioinas, J. Kannala, and E. Rahtu. Image-based localization using hourglass networks. pages 879–886, . URL http://openaccess.thecvf.com/content_ICCV_2017_workshops/w17/html/Melekhov_Image-Based_Localization_Using_ICCV_2017_paper.html.

[29] I. Melekhov, J. Ylioinas, J. Kannala, and E. Rahtu. Relative camera pose estimation using convolutional neural networks. In J. Blanc-Talon, R. Penne, W. Philips, D. Popescu, and P. Scheunders, editors, *Advanced Concepts for Intelligent Vision Systems*, volume 10617, pages 675–687. Springer International Publishing, . ISBN 978-3-319-70352-7 978-3-319-70353-4. doi: 10.1007/978-3-319-70353-4_57. URL http://link.springer.com/10.1007/978-3-319-70353-4_57.

[30] M. S. Müller, S. Urban, and B. Jutzi. SQUEEZEPOSENET: IMAGE BASED POSE REGRESSION WITH SMALL CONVOLUTIONAL NEURAL NETWORKS FOR REAL TIME UAS NAVIGATION. IV-2/W3:49–57. ISSN 2194-9050. doi: 10.5194/isprs-annals-IV-2-W3-49-2017. URL https://www.isprs-ann-photogramm-remote-sens-spatial-inf-sci.net/IV-2-W3/49/2017/.

[31] T. Naseer and W. Burgard. Deep regression for monocular camera-based 6-DoF global localization in outdoor environments. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1525–1530. IEEE. ISBN 978-1-5386-2682-5. doi: 10.1109/IROS.2017.8205957. URL http://ieeexplore.ieee.org/document/8205957/.

[32] H. Noh, S. Hong, and B. Han. Learning deconvolution network for semantic segmentation. pages 1520–1528. URL https://www.cv-foundation.org/openaccess/content_iccv_2015/html/Noh_Learning_Deconvolution_Network_ICCV_2015_paper.html.

[33] E. Parisotto, D. S. Chaplot, J. Zhang, and R. Salakhutdinov. Global pose estimation with an attention-based recurrent network. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 350–35009. IEEE. ISBN 978-1-5386-6100-0. doi: 10.1109/CVPRW.2018.00061. URL https://ieeexplore.ieee.org/document/8575522/.

[34] X. Peng, B. Sun, K. Ali, and K. Saenko. Learning deep object detectors from 3d models. URL http://arxiv.org/abs/1412.7122.

[35] N. Piasco, D. Sidibé, C. Demonceaux, and V. Gouet-Brunet. A survey on visual-based localization: On the benefit of heterogeneous data. 74:90 – 109. doi: 10.1016/j.patcog.2017.09.013. URL https://hal.archives-ouvertes.fr/hal-01744680.

[36] L. Pishchulin, A. Jain, M. Andriluka, T. Thormählen, and B. Schiele. Articulated people detection and pose estimation: Reshaping the future. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3178–3185. doi: 10.1109/CVPR.2012.6248052.

[37] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-CNN: Towards real-time object detection with region proposal networks. URL http://arxiv.org/abs/1506.01497.

[38] 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. URL http://arxiv.org/abs/1409.0575.

[39] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. URL http://arxiv.org/abs/1409.1556.

[40] H. Su, C. R. Qi, Y. Li, and L. Guibas. Render for CNN: Viewpoint estimation in images using CNNs trained with rendered 3d model views. URL `http://arxiv.org/abs/1505.05641`.

[41] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. pages 1–9. URL `https://www.cv-foundation.org/openaccess/content_cvpr_2015/html/Szegedy_Going_Deeper_With_2015_CVPR_paper.html`.

[42] A. Valada, N. Radwan, and W. Burgard. Deep auxiliary learning for visual localization and odometry. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6939–6946. IEEE, . ISBN 978-1-5386-3081-5. doi: 10.1109/ICRA.2018.8462979. URL `https://ieeexplore.ieee.org/document/8462979/`.

[43] A. Valada, N. Radwan, and W. Burgard. Incorporating semantic and geometric priors in deep pose regression. page 4, .

[44] G. Varol, J. Romero, X. Martin, N. Mahmood, M. J. Black, I. Laptev, and C. Schmid. Learning from synthetic humans. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4627–4635. IEEE. ISBN 978-1-5386-0457-1. doi: 10.1109/CVPR.2017.492. URL `http://ieeexplore.ieee.org/document/8099975/`.

[45] F. Walch, C. Hazirbas, L. Leal-Taixé, T. Sattler, S. Hilsenbeck, and D. Cremers. Image-based localization using LSTMs for structured feature correlation. URL `http://arxiv.org/abs/1611.07890`.

[46] S. Wang, R. Clark, H. Wen, and N. Trigoni. DeepVO: Towards end-to-end visual odometry with deep recurrent convolutional neural networks. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2043–2050. IEEE. ISBN 978-1-5090-4633-1. doi: 10.1109/ICRA.2017.7989236. URL `http://ieeexplore.ieee.org/document/7989236/`.

[47] J. Wu, L. Ma, and X. Hu. Delving deeper into convolutional neural networks for camera relocalization. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5644–5651. IEEE. ISBN 978-1-5090-4633-1. doi: 10.1109/ICRA.2017.7989663. URL `http://ieeexplore.ieee.org/document/7989663/`.