M. Etxeberria-Garcia, M. Labayen, M. Zamalloa and N. Arana-Arexolaleiba, "Application of Computer Vision and Deep Learning in the railway domain for autonomous train stop operation," *2020 IEEE/SICE International Symposium on System Integration (SII)*, Honolulu, HI, USA, 2020, pp. 943-948.

DOI: https://doi.org/10.1109/SII46433.2020.9026246

Application of Deep Learning in railway domain for train stop operation

Mikel Etxeberria-Garcia[1], Mikel Labayen[2], Maider Zamalloa[1] and Nestor Arana-Arexolaleiba[3]

Abstract— **The purpose of this paper is to present .. Design/methodology/approach - Short comings today... The paper proposes an approach consisting of ... Findings - Results from the ... shows that ... is capable of successfully ... Furthermore, the paper presents future research and development suggestions for XXX, which contributes to near-term industrial maturation and implementation. Originality/value - The paper presents a full-scale demonstration ... with particular focus on industrial utilization and application.**

I. INTRODUCTION

The application of Machine Learning has increased since the applicability of some of its techniques has improved. Deep Learning is one of the most growing techniques of Machine Learning. The good results that have achieved in recent researches and the increase of computational capacity have lead to a time where Deep Learning can be applied to a wide range of domains. From Medical technologies to the Internet of Things through its mayor domain, robotics. The power of Deep Learning in robotics lies in the potential it has of making a system that can learn. The robotics community has identified and summarized several applications for Deep Learning in robotics, such as, learning complex dynamics, control operations, advanced manipulation, object recognition or interpretation of human actions.

In this context, Deep Learning application has facilitated a development in the autonomous driving industry as one of the most important future business bets. Computer vision techniques, using Deep Learning, have helped to create machine-learning-based robots and cars that can predict and learn how to drive in various environments. The researches carried out recently on Intelligent Transportation Systems (ITS), Advanced Driving Assistance Systems (ADAS), intelligent infrastructures and autonomous driving have carried many benefits to the transportation industry. These technologies provide the vehicle its own decision-making capacity and the ability to interpret its environment, and consequently, enhance the control and signaling solutions. The irruption of Artificial Intelligence techniques in general and Deep Learning techniques, in particular, have allowed improving the perception capacity of these systems and the knowledge derived from the information perceived in the environment.

The railway domain is also transforming towards the ITS and ADAS industry. Nowadays, this sector is ready for the next steps involving itself in different research projects related to Computer Vision and Artificial Intelligence development.

In a fully autonomous train system all the operations involved in must be automatic, for example, visual odometry, people and obstacle detection-identification in railroads, operations such as train doors opening/closing, gauge control in platforms, coupling or as is presented in this work, accurate train stopping in train platforms.

This paper is divided into the following sections. Section II presents a use-case of a company related to the railway domain. Main approaches in train-robot localization using Deep Learning are explored in Section III, followed by the first use-case presentation in Section IV. Finally, some expected results are drawn in Section V.

II. PROBLEM DEFINITION

Communication-Based Train Control (CBTC) is a standard defined by the IEEE (IEEE 1474 [1]) which defines a set of performance and functional requirements for track and onboard equipment in order to enhance performance, availability, operations and the protection of the involved systems. A CBTC system could be defined as an automatic train control system where the track and onboard subsystems are continuously communicated. The main two functionalities covered by those subsystems are the Automatic Train Protection (ATP) and Automatic Train Operation (ATO). ATP subsystems monitor the train speed and position in order to guarantee a safe train operation. On the other hand, ATO subsystems are dedicated to the operations devoted to reaching a more autonomous and efficient train driving experience, such as, driving assistance tasks or automatic control of train brake and traction commands that aim to ensure that train speed is lower than the limit established by the ATP system [2].

Current CBTC systems, according to the standard IEC 62290-1, can be divided into pre-established Grades of Autonomy (GOA). The GOA of a train implementing any autonomous operation will have a value between 2 and 4: GOA2 for a semi-automated Train Operation, GOA3 for a driverless Train Operation and GOA4 for an unattended Train Operation. In GOA3 and GOA4 systems, as there is not a driver inside the train, an accurate train location system is required. Precise positioning systems can reach a higher grade of automation [3]. A train that implements GOA3 or GOA4 level can be considered as a robot that navigates

[1] Ikerlan Technology Research Centre, P. J. M. Arizmendiarrieta, 2 20500 Arrasate-Mondragn Gipuzkoa, Spain {mikel.etxeberria, mzamalloa}@ikerlan.es

[2] ??mikel.labayen@gmail.com

[3] Mondragon Unibertsitatea, Loramendi Kalea, 4 20500 Arrasate-Mondragn Gipuzkoa, Spain narana@mondragon.edu

This research is supported by Basque Government through BIKAINTEK2018 program and CAF Signalling. In collaboration between Ikerlan, CAF Signalling and Mondragon Unibertsitatea.

through a track in indoor and outdoor environments including underground stations. Therefore, it becomes essential to implement precise and reliable train localization subsystems. A GOA3 or GOA4 train must compute, among others, the braking curve or the train stopping location with precision.

Accurate train localization and platform-train doors alignment is essential for a safe passenger transfer train operation. Door equipped platform, which avoids human or undesirable objects fall in railway area, are more common than in the past. To align the doors and platform in a train stopping point requires a precise localization information. Nowadays, it is calculated using train speed data captured form different odometry sensors. These sensor errors are corrected time-to-time during train service using beacon information. However, in a stopping point the driver eyes and experience are still the key factors to align correctly the train with the platform area and to remove final localization error.

On the railway domain, most researches focus on other computer vision problems as object detection, although some of this approaches may be applied for localization purposes. The only research found using a monocular camera is DisNet [4] proposed by Haseeb *et al.*. They proposed a CNN to estimate the distance to previously detected objects by a monocular camera installed on a train. For the object detection part uses the standard YOLO [5] algorithm, also based on CNNs. However, most approaches in this domain are based on other type of sensors as stereo cameras [6], [7].

The main objective of this research is to explore the capability of Machine Learning techniques, particularly Deep Learning techniques, and computer vision for an accurate train stopping in fully autonomous train stop operation.

III. RELATED WORK

Lately, the capacity of Computer Vision to address some robotics problems has increased due to the rise of the application of Deep Learning algorithms. One of the solutions afforded efficiently by these strategies is the visual localization estimation. The sensors used in visual localization systems include monocular, stereo and RGB-D cameras. When choosing a sensor, the scale is essential. In visual problems some cameras are not able to calculate absolute scale and therefore, scale drifts [8]. Stereo cameras provide an immediate scale while requiring more calibration. RGB-D cameras provide color and depth information for each pixel in an image [9] but the cost is higher. In general, a lot of research interest has been focused on dense and semi-dense methods from a single camera [10]. In some researches, data fusion is also proposed as an approach to complement each sensor.

The use of Machine Learning approaches in Computer Vision problems has grown with the increase of computational resources and Deep Learning progress. This situation also comes from the promising results obtained by the application of deep approaches in computer vision, specifically with the use of Convolutional Neural Networks (CNN) on large-scale image classification (Krizhevsky *et al.* [11]). This work demonstrates the idea of the benefits of using CNNs on Computer Vision problems. Additionally, it has been shown that one of the potentials of CNN is their generalization ability in visual recognition tasks. A CNN trained for another purpose at first instance can be reused to solve another purpose without the need for full training phase again. Most systems use CNNs to find only local features or generate descriptors of discrete proposal regions [12]. Several works state that deep learning algorithms can model localization or depth solutions by regression [13], [14].

Three main techniques can be distinguished in visual localization problems: Visual Odometry (VO), Simultaneous Location and Mapping (SLAM) and Depth Estimation. Some of these techniques refer to the same problems, share viewpoints and in some cases can not be clearly differentiated.

- *Visual Odometry (VO).* Odometry can be defined as the use of data from motion sensors in order to estimate changes in position over time [15]. Visual odometry (VO) is a particular case of odometry where the position information is acquired through camera images [8]. The term Visual Odometry was first introduced by Niester *et al.* [16] proposing a method for estimating camera motion using RANSAC [17] outlier refinement method and tracking extracted features across all frames. Before that, feature matching was done just in consecutive frames. Later researches have shown that VO methods perform significantly better than wheel odometry in robotics while the cost of cameras is much lower compared to more accurate IMUs and LASER scanners [8]. This scenario raises the need for exploration of the applicability of VO in the railway domain and autonomous driving trains.
- *SLAM.* Simultaneous Localization and Mapping (SLAM) is a technique to reconstruct an unknown 3D environment. It has become a popular research topic as it is the base for autonomous robot navigation. Visual SLAM (vSLAM) is the field of SLAM comprised of methods that use visual information. Both vSLAM and VO can handle the same localization estimation problems and share a lot of components such as feature extractors. The main difference between both techniques is that VO centers on a relative part of the map while vSLAM uses the full context and global consistency is aimed [18].
- *Depth estimation.* Scene depth refers to the distance from the camera optical center to the object along to the optical axis [19]. The estimation of depth can contribute to localization, and in many approaches is used as a SLAM phase.

Since LeCun *et al.* [20] first explored the use of CNNs, several networks have been designed for a wide variety of problems. GoogLeNet/Inception was presented by Szegedy *et al.* [21] as an architecture of Deep CNNs, increasing the depth and width of previous architectures. Based on this inception concept, Chollet *et al.* [22] presented Xception CNN architecture based entirely on depthwise separable convolution layers. Then, VGGNet was defined by Simonyan *et al.* [23]. In this work, the authors evaluate the relevance of CNN depth on image classification tasks. Popular object detection network Fast R-CNN was presented by Girshick *et al.* [24] and later improved by Ren *et al.* [25] presenting Faster R-CNN. This two networks increased the training and testing speed using a region-based strategy. Based on these works, He *et al.* [26] presented ResNet, a residual learning framework to ease the training of deeper networks for image recognition, due to the difficulty of deeper neural networks to be trained. All of these networks have become the basis of most modern Deep Learning approaches.

Some recent works based on the ones previously mentioned, apply deep learning algorithms in VO solutions. They can estimate the pose directly from an input image without feature extraction or feature matching processes. In [27] Kendall *et al.* proposed PoseNet, a robust and real-time monocular re-localization system based on an end-to-end trained CNN. This approach was improved by adding a fundamental treatment of scene geometry introducing geometric loss functions [28]. Wang *et al.* [29] presented an approach that mixes CNN and RNN called Recurrent Convolutional Neural Network (RCNN). It takes the benefits of both networks, the feature extraction capabilities of the CNN and the sequential modeling from the RNN. In addition, Clark *et al.* [13] extended PoseNet with a RNN in order to exploit temporal dependencies and improve the monocular image sequence localization accuracy. Some approaches that extend the PoseNet system have been presented, ie. relative ego-motion [30]. Later, in [31] Xiang *et al.* introduced PoseCNN, a CNN for 6D object pose estimation. PoseCNN localizes an object center in the image and predicts its distance from the camera.

From all the explored techniques three of them has been selected as the most interesting and relevant for our particular use case:

- Disnet. It is the only approach based on the same domain and is based on CNNs. Uses YOLO to detect objects and then DisNet network to regress the distance to those objects.
- PoseCNN. Detects the center of a known object and estimates the distance from the camera to regress the pose of that object using a CNN.
- DeepVO. Introduces Recurrent Networks to the localization problem and takes advantage of the input videos as it infers poses of objects directly from a sequence of images.

The application of these techniques is foreseen in a real-world use case from the railway environment as there are few applications of deep learning approaches in this domain due to strict railway regulation. The main goal is to explore the applicability of Deep Learning for Visual Odometry, SLAM and Depth estimation in a different domain.

IV. USE CASE: TRAIN STOP OPERATION

A. *Use case definition*

This section defines an Autonomous Urban Train use-case where artificial intelligence and high-performance computational capabilities are used to increase the dependability and the safety of the system. The objective is to apply Computer Vision and Deep Learning techniques to improve different autonomous train operation functionalities as precision stop, rolling stock coupling operation or person and obstacle detection-identification in railroads.

The selected use-case is the automatic accurate stop at door equipped platforms aligning the vehicle and platform doors. The goal is to perform precise localization inside platform area using visual patterns detection, identification and tracking in order to reach accurate stopping point and managing automatic train operation (traction and brake commands, ATO functionality). A contribution is foreseen to the automatic train operation system, adding the visual localization estimation information to the usual trains odometry data calculations based on radars and encoders.

In the current train localization system, beacon positions are known by trackside equipment and may be known by train if previously announced. From beacon to beacon, a localization error is accumulated that is proportional to traveled distance. Each time the train crosses a beacon, the localization and accuracy is reset. With the combination of wheel odometry data, given by radars and encoders, and Visual Odometry (VO) data, provided by our proposed approaches, an improvement on the precision of the stop is foreseen, where the localization error must be lower than the current error given by beacon-based train localization system.

The main idea of our approach is to detect a pattern that is always placed on the platform that will help us to locate the train through Deep Neuronal Networks (DNN). These patterns usually are used by train drivers to know the stopping position of the train and have a regular form and color. One example is shown in figure 1

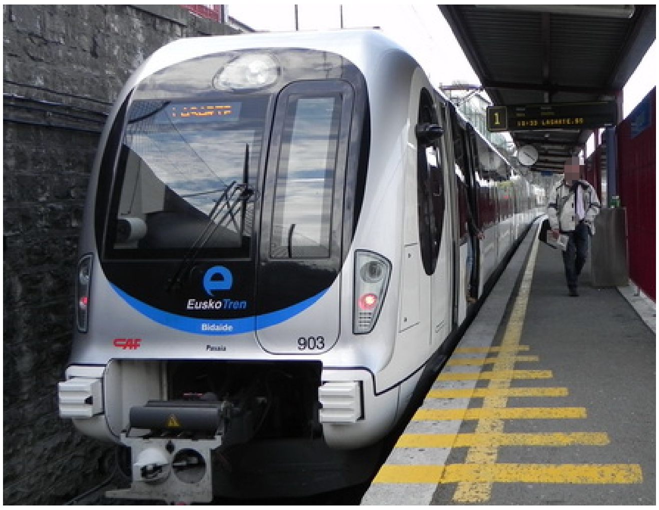

Fig. 1. Signaling patterns are shown at the end of the platforms

B. *Architecture*

The architecture of the designed application for this use-case is shown in figure 2. The videos are captured using a camera that transfers the images to a capturer, which has two workflows. First, transfers the videos to a database (DB) that will be used to pre-process the videos and train a DNN. The training process of the DNN will gave a model that will be used later for real time processing. Secondly, the capturer passes the streaming of frames to the previously trained DNN that will output the desired result. Depending on the selected approach, the pre-process done to the input videos, the structure of the DNN and the output will be different. Usually the pre-process phase is done using Computer Vision techniques without Machine Learning.

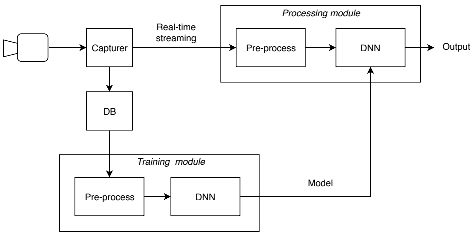

Fig. 2. Architecture of the designed application for train localization using Computer Vision and Deep Neural Networks (DNN)

C. Datasets and data collection

Deep Learning approaches require large amounts of data for training. This data can be collected from different sources: collected for this particular research, using simulated environments and from existing datasets. The first option is to use data from the standard datasets created by other institutions or researches previously. Having a properly labeled ground truth in these datasets is essential as they are the base for training DNNs and evaluating performances. Depending on the selected approach the required ground truth data is not the same. After an analysis of the most used databases for visual localization researches, we have found that only one database (Norland [32]) covers the railway domain and it does not match our need for indoor images. A brief summary of database analysis is shown in table I. For each database, the used sensors, the domain it belongs to, if they give pose or/and depth information and if it is an indoor or an outdoor research is addressed.

To collect data, an appropriate environment is required, where a camera can be used to take images from the front of the train to the track. The advantage of this system is that database can be designed to the particular use-case, but it requires a lot of time of recording and permissions for experimentation in the track. To overcome this problem, simulated environments can be used, where no real railways are involved. The drawback of simulated environments is that we can not assure that an algorithm trained and validated in a simulated environment will give the same results in a real world scenario.

Although the first tests will be in a simulated 3D environment, we will eventually have access to real trains to create a database to test the application of the selected approaches. The data base collected to test the functionality will contain enough variety in scenarios taking into account unfavorable light conditions, pattern shape/color degradation due to passing of time and partial occlusions (hidden patterns because of people or object presence).

V. EXPECTED RESULTS AND CONCLUSIONS

The main goal of the research is to explore the application of deep learning techniques in railway domain as there is not much research in this domain, although we can make use of approaches from other domain where the research has increased lately, i.e. robotics or other autonomous vehicles. This article presents a use-case for the application of deep learning techniques in railway domain for precise localization in an indoor environment. It also shows the main architecture of the system build to solve the presented use-case. Finally, shows the high data quantity required by deep learning solutions, and the need of a dataset for our approach that has not been found in current most used datasets.

Railway operators are interested in market more accessible and flexible solutions aligned with social sustainability and mobility concerns fulfilling urban operators need. If urban vehicles (metro) gain autonomy, system development costs are reduced (install and maintenance costs) and operation flexibility is gained. Giving to vehicles autonomy and decision-making capabilities complements the information already received from railroad signaling modes as they can observe and interpret the environment in an independent manner.

As expected results, the use-case application must be able to perform accurate automatic stop at door equipped platforms, aligning the vehicle and platform for correct passenger transfer. For that, a accurate localization is expected from selected approaches. According to the accuracy requirements, alignment error should be lower than 5 cm at 99,9% of times (measure errors must to be taken into account).

REFERENCES

[1] "Ieee standard for communications-based train control (cbtc) performance and functional requirements," *IEEE Std 1474.1-2004 (Revision of IEEE Std 1474.1-1999)*, pp. 1–45, 2004.

[2] M. Malvezzi, B. Allotta, and M. Rinchi, "Odometric estimation for automatic train protection and control systems," *Vehicle System Dynamics*, vol. 49, no. 5, pp. 723–739, 2011.

[3] T. Albrecht, K. Lüddecke, and J. Zimmermann, "A precise and reliable train positioning system and its use for automation of train operation," *IEEE ICIRT 2013 - Proceedings: IEEE International Conference on Intelligent Rail Transportation*, pp. 134–139, 2013.

[4] M. A. Haseeb, J. Guan, D. Ristić-Durrant, and A. Gräser, "Disnet: A novel method for distance estimation from monocular camera," 2012.

[5] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[6] J. Weichselbaum, C. Zinner, O. Gebauer, and W. Pree, "Accurate 3d-vision-based obstacle detection for an autonomous train," *Computers in Industry*, vol. 64, no. 9, pp. 1209–1220, 2013.

[7] N. Fakhfakh, L. Khoudour, E.-M. El-Koursi, J.-L. Bruyelle, A. Dufaux, and J. Jacot, "Background subtraction and 3d localization of moving and stationary obstacles at level crossings," in *2010 2nd International Conference on Image Processing Theory, Tools and Applications*. IEEE, 2010, pp. 72–78.

[8] K. Yousif, A. Bab-Hadiashar, and R. Hoseinnezhad, "An Overview to Visual Odometry and Visual SLAM: Applications to Mobile Robotics," *Intelligent Industrial Systems*, vol. 1, no. 4, pp. 289–311, 2015.

Dataset	Domain	Sensors	Pose	Depth	Indoor/Outdoor
SUN3D [33]	Robot	RGB-D camera	X	X	I
TUM-LSI [34]	Robot	RGB-D camera	X		I
NavVis [34]	Robot	Camera	X		I
Cambridge [27]	Robot	Smartphone	X		O
7-scenes [35]	Urban localization	RGB-D camera	X		I
BigSFM [36][37]	Urban localization	Camera	X		O
MIT DATA [38]	Robot	LIDAR, Stereo camera, Odometry	X		I
KITTI [39][40]	Car	Laser, Stereo camera, GPS	X	X	O
Nordland [32]	**Train**	Camera, GPS/INS	X		O
Oxford RobotCar [41]	Car	Cameras, LIDAR, GPS/INS	X		O
EuroC/MAV [42]	Micro Aerial Vehicle	Stereo camera, Laser, IMU	X	X	I
The Wean Hall [43]	Robot	Stereo camera, Laser, IMU	X	X	I
Ford Campus [44]	Car	Camera, LIDAR, IMU	X		O
RGB-D SLAM [45]	Robot	RGB-D camera, Accelerometer	X	X	I
GMU Kitchen [46]	3D reconstruction	RGB-D camera	X	X	I
NYUD/NYUD2 [47][48]	3D reconstruction	RGB-D camera		X	I
ETH3D [49]	3D reconstruction	Camera, Laser, IMU	X	X	I/O
Make3D [50]	3D reconstruction	Camera, Laser	X	X	I/O
MPI-Sintel [51]	Movie	(Digital)		X	I/O

TABLE I

MOST USED DATASETS FOR LOCALIZATION

[9] S. Poddar, R. Kottath, and V. Karar, "Evolution of Visual Odometry Techniques," 2018. [Online]. Available: http://arxiv.org/abs/1804.11142

[10] K. Tateno, F. Tombari, I. Laina, and N. Navab, "CNN-SLAM : Real-time dense monocular SLAM with learned depth prediction," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6243–6252.

[11] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[12] D. Eigen and R. Fergus, "Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2650–2658.

[13] R. Clark, S. Wang, A. Markham, N. Trigoni, and H. Wen, "VidLoc : A Deep Spatio-Temporal Model for 6-DoF Video-Clip Relocalization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6856–6864.

[14] H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao, "Deep Ordinal Regression Network for Monocular Depth Estimation," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 2002–2011, 2018.

[15] J. Otegui, A. Bahillo, I. Lopetegi, and L. E. Diez, "A Survey of Train Positioning Solutions," *IEEE Sensors Journal*, vol. 17, no. 20, pp. 6788–6797, 2017.

[16] D. Nistér, O. Naroditsky, and J. Bergen, "Visual Odometry," in *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, 2004, p. 1.

[17] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[18] T. Taketomi, H. Uchiyama, and S. Ikeda, "Visual slam algorithms: A survey from 2010 to 2016," *IPSJ Transactions on Computer Vision and Applications*, vol. 9, no. 1, p. 16, 2017.

[19] L. He, G. Wang, and Z. Hu, "Learning depth from single images with deep neural network embedding focal length," *IEEE Transactions on Image Processing*, vol. 27, no. 9, pp. 4676–4689, 2018.

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

[21] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[22] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1251–1258.

[23] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[24] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.

[25] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," *Advances in neural information processing systems*, pp. 91–99, 2015.

[26] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[27] 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.

[28] A. Kendall and R. Cipolla, "Geometric loss functions for camera pose regression with deep learning," *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, vol. 2017-Janua, pp. 6555–6564, 2017.

[29] 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)*. IEEE, 2017, pp. 2043–2050.

[30] 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.

[31] Y. Xiang, T. Schmidt, V. Narayanan, and D. Fox, "Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes," *arXiv preprint arXiv:1711.00199*, 2017.

[32] D. Olid, J. M. Fcil, and J. Civera, "Single-view place recognition under seasonal changes," in *PPNIV Workshop at IROS 2018*, 2018.

[33] J. Xiao, A. Owens, and A. Torralba, "Sun3d: A database of big spaces reconstructed using sfm and object labels," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 1625–1632.

[34] F. Walch, C. H. L. Leal-taix, T. Sattler, and S. H. D. Cremers, "Image-based localization using LSTMs for structured feature correlation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 627–637.

[35] B. Glocker, S. Izadi, J. Shotton, and A. Criminisi, "Real-time rgb-d camera relocalization," in *2013 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE, 2013, pp. 173–179.

[36] Y. Li, N. Snavely, and D. P. Huttenlocher, "Location recognition using prioritized feature matching," in *European conference on computer vision*. Springer, 2010, pp. 791–804.

[37] D. Crandall, A. Owens, N. Snavely, and D. Huttenlocher, "Discrete-continuous optimization for large-scale structure from motion," in *CVPR 2011*. IEEE, 2011, pp. 3001–3008.

[38] M. Fallon, H. Johannsson, M. Kaess, and J. J. Leonard, "The mit stata center dataset," *The International Journal of Robotics Research*, vol. 32, no. 14, pp. 1695–1699, 2013.

[39] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 3354–3361.

[40] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1231–1237, 2013.

[41] W. Maddern, G. Pascoe, C. Linegar, and P. Newman, "1 year, 1000 km: The oxford robotcar dataset," *The International Journal of Robotics Research*, vol. 36, no. 1, pp. 3–15, 2017.

[42] M. Burri, J. Nikolic, P. Gohl, T. Schneider, J. Rehder, S. Omari, M. W. Achtelik, and R. Siegwart, "The euroc micro aerial vehicle datasets," *The International Journal of Robotics Research*, vol. 35, no. 10, pp. 1157–1163, 2016.

[43] H. Alismail, B. Browning, and M. B. Dias, "Evaluating pose estimation methods for stereo visual odometry on robots," in *the 11th Intl Conf. on Intelligent Autonomous Systems (IAS-11)*, vol. 3, 2010, p. 2.

[44] G. Pandey, J. R. McBride, and R. M. Eustice, "Ford campus vision and lidar data set," *The International Journal of Robotics Research*, vol. 30, no. 13, pp. 1543–1552, 2011.

[45] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers, "A benchmark for the evaluation of rgb-d slam systems," in *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2012, pp. 573–580.

[46] G. Georgakis, M. A. Reza, A. Mousavian, P.-H. Le, and J. Košecká, "Multiview rgb-d dataset for object instance detection," in *2016 Fourth International Conference on 3D Vision (3DV)*. IEEE, 2016, pp. 426–434.

[47] N. Silberman and R. Fergus, "Indoor scene segmentation using a structured light sensor," in *2011 IEEE international conference on computer vision workshops (ICCV workshops)*. IEEE, 2011, pp. 601–608.

[48] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from rgbd images," in *European Conference on Computer Vision*. Springer, 2012, pp. 746–760.

[49] T. Schops, J. L. Schonberger, S. Galliani, T. Sattler, K. Schindler, M. Pollefeys, and A. Geiger, "A multi-view stereo benchmark with high-resolution images and multi-camera videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3260–3269.

[50] A. Saxena, S. H. Chung, and A. Ng, "Learning Depth from Single Monocular Images," *Advances in Neural Information Processing Systems 18*, pp. 1161–1168, 2006. [Online]. Available: http://books.nips.cc/papers/files/nips18/NIPS2005_0684.pdf

[51] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black, "A naturalistic open source movie for optical flow evaluation," in *European conference on computer vision*. Springer, 2012, pp. 611–625.