Open Access
Online Journal of
Robotics & Automation Technology
Iris Publishers

Research Article

CNN-Derived Local Features for Condition-Invariant Robot Localization with a Single RGB Sensor

Loukas Bampis* and Antonios Gasteratos

Department of Production and Management Engineering, Democritus University of Thrace, Greece

***Corresponding author:** Loukas Bampis, Department of Production and Management Engineering, Democritus University of Thrace, Xanthi, Greece.

Received Date: February 23, 2021

Published Date: March 04, 2021

Abstract

A conditio sine qua non for any autonomous system refers to its ability to localize itself into a known environment. Towards this end, camera sensors are typically deployed for measuring the relative transformation between the mobile platform and a pre-mapped scene due to their low cost and substantial accuracy. This paper deals with the task of robot localization under different environmental conditions using a single RGB sensor. To achieve this, we diverge from conventional approaches that detect local points of interest using hand-crafted sets of rules, and we utilize the cognitive properties of modern deep learning models for feature detection and description. The proposed architecture is evaluated and compared against other traditional techniques under a wide range of environmental conditions that significantly alter the view of a previously recorded area.

Keywords: Autonomous Systems; Robot Localization; Deep Learning; Local Features

Introduction

Navigation, obstacle avoidance, manipulation are some of the most representative functionalities that an autonomous agent needs to carry out in order to effectively perform real-world activities. Yet, none of the above would be possible if the respective platform's location within the environment could not be retrieved. Therefore, localization techniques have been developed to identify the agent's position and orientation in the three-dimensional (3D) space. Absolute positioning sensors, like Global Navigation Satellite System (GNSS) [1] or WiFi signal receivers [2], may seem straightforward solutions for addressing such a task; however, they fail to provide adequate accuracy and can only be applied in specific operational conditions. Instead, a wide variety of exteroceptive sensors have been adopted by the related literature, such as sonars [3], laser rangefinders [4], stereoscopic cameras [5], etc., to deduce the platform's relative transformation from a previously mapped scene. During the last decade, driven by the market's commitment to producing low-cost systems, the scientific community's attention has focused on monocular RGB camera solutions due to their lower weight and power consumption needs [6,7].

In the typical case, given a pre-recorded map of the environment, single-image localization is achieved by detecting local key-points in the captured image and comparing them with the ones of a known map (Figure 1).

Local key-points represent prominent points [8] or blobs [9] in a given frame that can be effectively re-detected in the same manner regardless of the camera's viewing angle. Each detected key-point is assigned with a description vector that captures illumination differences among pixels within a patch centered around its origin. Such descriptors are distinctive for each local feature. Thus, calculating distance metrics between them and identifying nearest neighboring ones can effectively yield key-point matches from different camera measurements of the same entity in the environment. Associating a sufficient number of local key-points allows for the utilization of camera displacement models (e.g., fundamental matrix, Perspective-n-Point) that can compute the relative transformation between a query frame and the already known map's structure [5,10].

Despite their effectiveness in identifying point matches between recordings of the same area under different viewing angles

(rotation- and scale-invariance), hand-crafted local features perform poorly [11] under extreme environmental changes that induce visual differences on the observed scene Figure 1-top. Such effects may refer to illumination variations between different periods of the day (morning-night) or weather conditions (sunny-rainy). Due to recent advancements in the field of deep learning, a variety of feature extraction algorithms have been proposed that base their functionality on Convolutional Neural Networks (CNN). By [12], or even pre-defined models for object detection [13], feature vectors are extracted from the total of each captured frame. Those vectors are treated as global descriptors for measuring similarity between the camera recordings' content. However, they remain unsuitable for the task of precise pose localization since they do not provide point-to-point associations, and thus, they cannot be used to retrieve the platform's position. Most recently, though, CNN architectures capable of detecting and describing local feature points have been reported [14,15], which are yet to be evaluated under the task of condition-invariant robot localization. As a final note, recently proposed CNN-based models have been developed to directly estimate the relative transformation between input frames [16] offering promising accuracy results.

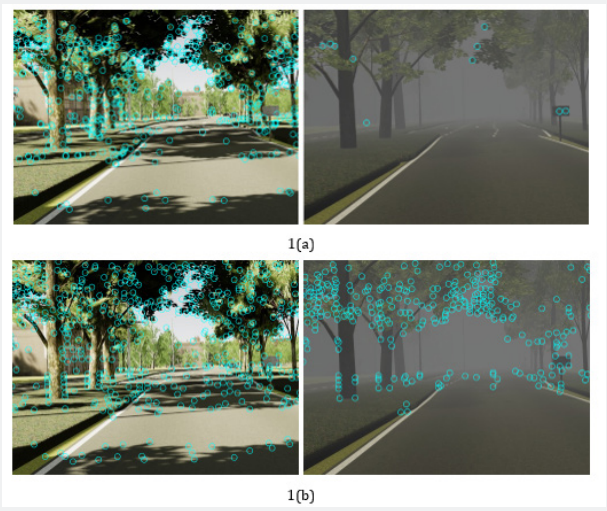

Figure 1: Motivating the necessity of condition-invariant local features for RGB camera-based localization. (Top) Key-point correspondences from hand-crafted features are not sufficient. (Bottom) Using a CNN-based architecture yields more detections allowing the computation of relative transformations between the two environmental conditions.

In this paper, we present a 6 Degrees of Freedom (6-DoF) localization mechanism using a single RGB sensor, which is capable of detecting feature points and producing the corresponding description vectors with invariance over the changing environmental conditions Figure 1-bottom. This allows for an autonomous platform to localize and operate, even if it known map was recorded under different conditions than the time of deployment, extending the potentials for applications like pose estimation, city-scale landmark detection, teach-and-repeat, or simultaneous localization and mapping. Section 2 presents our approach for accurate pose localization. We begin by describing the primary mechanism for computing relative image transformations, and subsequently, the model for producing CNN-derived Local feat Ur Es (CLUEs for sort) under different environmental conditions. Such a technique differs from the holistic proposal of assigning the whole localization procedure onto a CNN (such as in [16]). Instead, deep learning is only used for computing data associations, allowing for the utilization of well-established mathematical models and localization techniques from the robotics vision community [7,17]. In Section 3, CLUEs are evaluated, and their performance is compared against traditional hand-crafted features. Finally, Section 4 draws our conclusions and lists our plans for future work.

Approach

Single image-based localization

Our problem formulation dictates that an autonomous agent's operational environment has already been mapped during an exploration mission. When the robot attempts to interact with a previously recorded scene Id, localization of a new camera measurement I_q is achieved by computing its relative position with respect to the known one. To that end, we begin by following a standard procedure for feature matching [9]. Firstly, local key-points and descriptors are produced from each instance, resulting in sets P_q - P_d and D_q - D_d, respectively. Then, for each description vector $d_q \in D_q$, we identify the two nearest neighboring ones in D_d, viz, d_d^1 and d_d^2. Furthermore, in cases where multiple dq instances share a common nearest neighbor in D_d, we retain the one showing the lowest distance, thus ensuring unique associations. The above procedure allows to compute a normalized distance metric [9]:

$$S = \left\| d_q - d_d^1 \right\|_2 = \left\| d_q - d_d^2 \right\|_2, \qquad (1)$$

where the term $\|...\|_2$ denotes the L2 norm. Thus, feature matches are achieved by comparing distance S with a predefined threshold value th. With the above matches identified, we proceed by computing the camera's transformation between frames I_q and I_d. More specifically, we deploy the 8-Point Algorithm under a Random Sample Consensus (RANSAC) [18] scheme that effectively estimates the fundamental matrix by utilizing P_q and P_d coordinates of the matched features. As evident from the above, the selected local features need to offer repeatability and distinctiveness, even if the scene's view has been significantly altered due to different environmental conditions (e.g., day-night, sunny-rainy). Within the scope of this paper, repeatability essentially translates into the method's capacity for re-detecting each point from different camera instances, while distinctiveness evaluates the descriptors' performance for providing accurate point matches.

CNN-derived local features

In order to effectively meet the above conditions, we make use of Super Point [14], which represents a highly acknowledged CNN-based method for key-point detection and description. The network's first layers correspond to a VGG-based [19] encoder to reduce the input image's dimensions. The resulting tensor is fed into two different decoders, namely:

- An interest-point detection one, responsible for evaluating each key-point's prominence, and
- A description one, which is trained to produce feature vectors that maintain low distance between matching patches. The model's weights are all trained simultaneously under a joint loss function that accumulates the individual decoders' losses.

Super Point corresponds to a self-supervised architecture since it requires neither key-point annotations nor matching knowledge during training. Instead, the procedure begins by producing a pseudo-ground-truth image dataset containing synthetic shapes with homographic warps of simplified 2D geometries, such as lines, triangles, quadrilaterals, and ellipses. As proposed by the authors, we use the resulting model to generate key-points P_i on another training dataset depicting the autonomous agent's operational environment conditions. Images from this dataset are also homographically warped by a set of H transformations that successfully simulate multiple observations of the same area under varying viewpoint angles. To further simulate the effect of different environmental conditions, we also perform illumination changes on the warped instances by applying random brightness and contrast adjustments. Since matrices H are known, the detected P_i can be accurately projected on the warped images to produce the required key-point matching ground-truth in an unsupervised manner. Within the scope of this work, we perform the aforementioned homographic adaptation twice to better generalize the synthetic dataset's properties. With the above model trained, a set of CLUEs can be retrieved for any given input camera measurement I.

Experimental Results

To effectively evaluate the performance of a localization approach, we need to assess the repeatability and distinctiveness of a given local feature technique under different environmental conditions, as well as their sufficiency for computing fundamental matrices, as described in Section 2.1. Unfortunately, there are no currently available datasets for robot localization that include ground-truth information for point-to-point associations between sequences recorded from the same area under different environmental conditions. An alternative way for producing such data is to select a dataset that provides camera localization (6-DoF) and depth ground-truth for each of the images' pixels. With the above information at hand, key-points $\left({}^{I_1}P = \left[{}^{I_1}p_x, {}^{I_1}p_y \right]^T \right)$ can be detected from the camera measurements' undistorted equivalents at a particular condition (C_1), then projected back to the 3D world ${}^{W}P$, and finally, re-projected on the sequence frames $\left({}^{I_1}P = \left[{}^{I_2}p_x, {}^{I_2}p_y \right]^T \right)$ of a different condition (C_2). The mathematical formulation of the above steps is governed by the following set of equations:

$$ {}^{C_1}P = K_1^{-1} \begin{bmatrix} {}^{I_1}p_x \\ {}^{I_1}p_y \\ 1 \end{bmatrix} {}^{I_1}p_d$$

$$\begin{bmatrix} {}^{C_2}P \\ 1 \end{bmatrix} = {}^{C_2}_{W}T \begin{bmatrix} {}^{W}P \\ 1 \end{bmatrix} = {}^{C_2}_{W}T \bullet {}^{W}_{C_2}T \quad (2)$$

$$s \begin{bmatrix} {}^{I_2}p_x \\ {}^{I_2}p_y \\ 1 \end{bmatrix} = K_2 {}^{C_2}P$$

In the above, K_1 and K_2 denote the cameras' intrinsic matrices, ${}^{I_1}p_d$ is the depth value for the specific point ${}^{I_1}P$ as provided by the dataset's ground-truth, ${}^{C_1}P$ and ${}^{C_2}P$ are the point's 3D positions with respect to the cameras C_1 and C_2, respectively, while ${}^{W}_{C_1}T$ and ${}^{C_2}_{W}T$ are the transformation matrices obtained by the 6-DoF ground-truth position of each camera with respect to the world frame of reference. For the above procedure to offer the level of accuracy that ${}^{I_1}P$ to ${}^{I_2}P$ ground-truth associations require, the localization and depth data need to be eminently precise. For instance, a small error in the cameras' orientation may result in a completely false point projection between the two camera sensors (Figure 2,3).

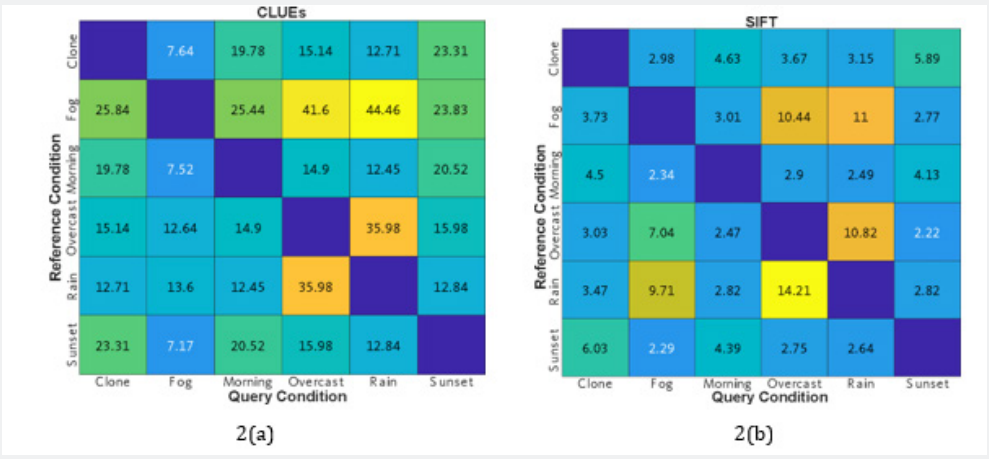

Figure 2: Average key-point detection repeatability percentage (%) for each condition pair, by CLUEs and SIFT features.

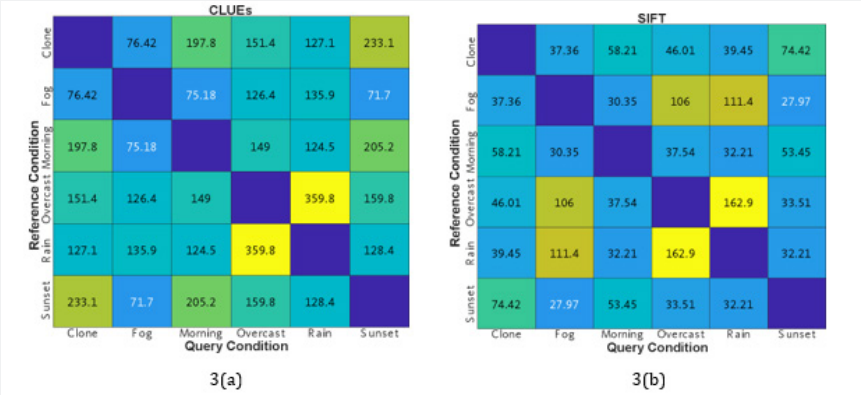

Figure 3: Average number of commonly detected points for each condition pair, by CLUEs and SIFT features.

To that end, until initiatives like the 4Seasons dataset [20] become publicly available, we make use of the vKITTI2 [21] dataset -the most recent version of Virtual KITTI [22]- which meets all the above conditions since it is captured under a virtual environment with all the localization data defined in an absolute manner. vKITTI2 contains 5 sequences, each of which being recorded under 6 different environmental conditions, namely clone[2], fog, morning, overcast, rain, and sunset. To evaluate the CNN architecture's performance, we make use of sequence 02, while the rest are utilized for training the model. Note that since vKITTI2 additionally offers per-pixel classification labels, we are also able to exclude dynamic classes, such as cars, from the training procedure. Finally, sequence02 is also exploited to assess the performance of SIFT [9], one of the most acknowledged hand-crafted local feature techniques, and obtain a comparative baseline.

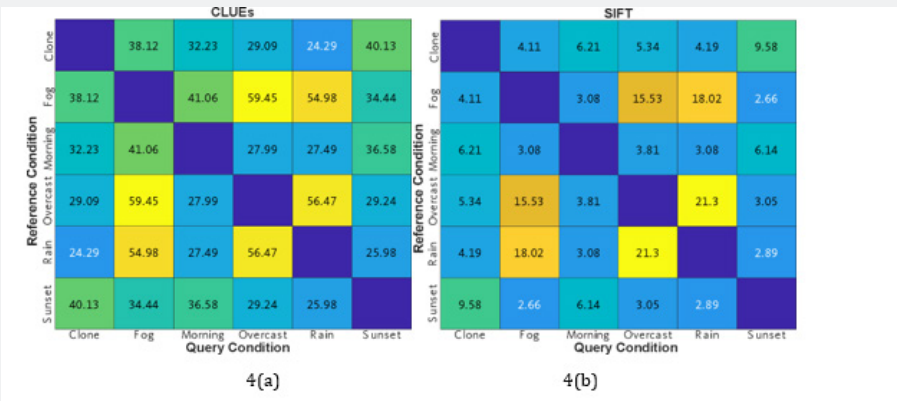

Figure 4: Average accurate descriptor matching percentage (%) for each condition pair, by CLUEs and SIFT features.

In this subsection, we evaluate the performance of CLUEs for detecting key-points between different environmental conditions in a repeatable manner. First, we deploy feature detection on every image of all the available conditions in sequence02. Then, the number of commonly detected key-points between all image pairs can be computed using the set of Equations 2. These metrics are normalized by the total number of detections per instance, resulting in the method's repeatability performance. Figure 2 shows the results we collected by averaging the repeatability metrics among all image-members for each condition. To obtain a better understanding of each method's behavior, we also present the average number of commonly detected feature points in Figure 3, which corresponds to an un-normalized equivalent of the results in Figure 2. As it can be seen, CLUEs outperform the detection performance of SIFT in every case, while also offering a sufficient number of repeatable detections to estimate the cameras' relative transformation (see Section 2.1).

Descriptors matching

When an autonomous agent operates, there is no information for associating the measurements' detected points with the ones of the reference sequence Equation 1. For this reason, feature descriptors need to preserve distinctiveness in order for them to be appropriately matched. To that end, Figure 4 depicts the average percentage of accurate descriptor matches between the different conditions in sequence02. Once more, CLUEs continuously outperform the results of SIFT descriptors, highlighting the superiority of a trainable architecture. It is noteworthy that the average normalized distance S between nearest neighboring CLUEs descriptors was recorded at 0.848, with the correctly and falsely associated ones showing 0.784 and 0.893, respectively. For the case of SIFT, the corresponding average S metrics were 0.259 for the total, 0.222 for the correctly matched, and 0.262 for the mismatched descriptor pairs. This essentially shows that CLUEs are generally closer in their respective descriptors' space than SIFT; however, they offer a sufficient value range for selecting a threshold and successfully distinguishing between true and false matches. For the rest of our experimentation, we choose generic midrange the values, viz. 0.839 for CLUEs and 0.242 for SIFT.

Localization performance

To demonstrate the localization performance of CLUEs, we compare it against the corresponding results of SIFT local features by computing relative transformations between different environmental conditions. Figure 5 presents the number of accurately estimated fundamental matrices for each condition pair. Note that sequence02 contains a total of 233 camera measurements per condition. As seen, SIFT managed to achieve perfect localization performance between cases that affect the environment's view in a similar manner (e.g., overcast-rain); however, the dominance of CLUEs is evident in all experiments.

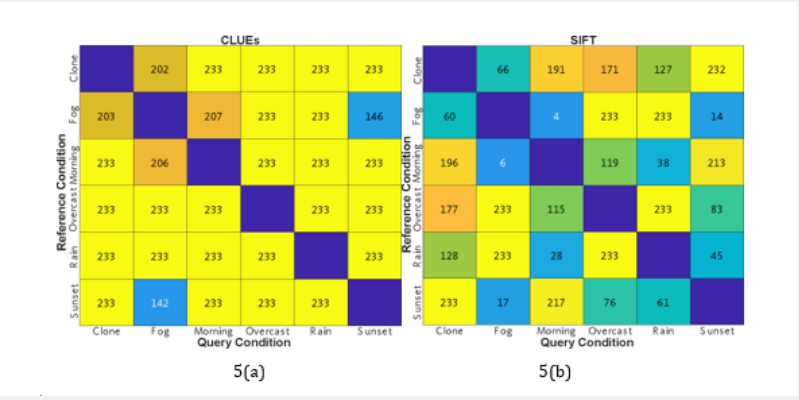

Figure 5: Number of successfully localized camera measurements by means of CLUEs and SIFT features.

Conclusion

In this paper, we presented a localization technique capable of operating under different environmental conditions using a single RGB sensor. To that end, the cognitive properties of Super Point were exploited and extended to images with significantly different appearances. The performance results were compared against an analogous technique for local feature extraction, which is based on hand-crafted rules for the key-points' detection and description. The CNN-based approach's superior localization prove that autonomy can be broadened with little-to-non cost since the operational environment needs to be mapped only once, during a particular period of the day or weather condition. Future work includes the our approach's evaluation on real-world datasets [20], once they become publicly available, and the adaptation of more prominent network layers [23,24].

Acknowledgment

This research is co-financed by Greece and the European Union (European Social Fund- ESF) through the Operational Programme

«Human Resources Development, Education and Lifelong Learning» in the context of the project "Reinforcement of Postdoctoral Researchers - 2nd Cycle" (MIS-5033021), implemented by the State Scholarships Foundation (IKY).

Conflict of Interest

Author declare no conflict of interest.

References

1. G Reina, A Vargas, K Nagatani, K Yoshida (2007) Adaptive Kalman Filtering for GPS-Based Mobile Robot Localization. In: Proc. IEEE Int Workshop Safety, Security and Rescue Robotics): 1-6.
2. W Xue, X Hua, Q Li, K Yu, W Qiu (2018) Improved Neighboring Reference Points Selection Method for Wi-Fi Based Indoor Localization. IEEE Sensors 2(2): 1-4.
3. H Liu, F Sun, B Fang, X Zhang (2016) Robotic Room-Level Localization Using Multiple Sets of Sonar Measurements. IEEE Trans Instrum Meas 66(1): 2-13.
4. F Browne, ST Padgett (2018) Novel Method of Determining Vehicle Cartesian Location Using Dual Active Optical Beacons and a Rotating Photosensor. IEEE Sensors Lett 2(4): 1-4.
5. R Mur Artal, JD Tardós (2017) ORB-SLAM2: An Open-Source SLAM System for Monocular, Stereo, and RGB-D Cameras. IEEE Trans Robot 33(5): 1255-1262.
6. White, DK Borah, W Tang (2019) Robust Optical Spatial Localization Using a Single Image Sensor. IEEE Sensors Lett 3(6): 1-4.
7. R Mur Artal, JMM Montiel, JD Tardos (2015) ORB-SLAM: A Versatile and Accurate Monocular SLAM System. IEEE Trans Robot 31(5): 1147-1163.
8. E Rosten, T Drummond (2006) Machine learning for high-speed corner detection. European Conference on Computer Vision, pp. 430-443.
9. DG Lowe (2004) Distinctive Image Features from Scale-Invariant Keypoints. International Journal of Computer Vision 60(2): 91-110.
10. L Bampis, A Gasteratos (2019) Revisiting the Bag-of-Visual-Words Model: A Hierarchical Localization Architecture for Mobile Systems. Robotics and Autonomous Systems (113): 104-119.
11. C Valgren, AJ Lilienthal (2010) SIFT, SURF & Seasons: Appearance-Based Long-Term Localization in Outdoor Environments. Robotics and Autonomous Systems 58(2): 149-156.
12. R Arandjelovic, P Gronat, A Torii, T Pajdla, J Sivic (2016) Net VLAD: CNN Architecture for Weakly Supervised Place Recognition. IEEE Conf Comput Vision and Pattern Recognition, pp. 5297-5307.
13. N Sünderhauf, S Shirazi, F Dayoub, B Upcroft, MJ Milford (2015) On the Performance of Conv Net Features for Place Recognition. IEEE/ RSJ Int Conf Intelligent Robots and Syst, pp. 4297-4304.
14. D DeTone, T Malisiewicz, A Rabinovich (2018) Super Point: Self-Supervised Interest Point Detection and Description. IEEE Conf Comput Vision and Pattern Recognition, pp. 224-236.
15. M Yi, E Trulls, V Lepetit, P Fua (2016) LIFT: Learned Invariant Feature Transform. European Conf Comput Vision, pp. 467-483.
16. G Costante, M Mancini (2020) Uncertainty Estimation for Data-Driven Visual Odometry. IEEE Trans Robot 36(6): 1738-1757.
17. Engel, T Schöps, D Cremers (2014) LSD-SLAM: Large-scale direct monocular SLAM. European Conf Comput Vision, pp. 834-849.
18. D Gálvez-López, JD Tardós (2012) Bags of Binary Words for Fast Place Recognition in Image Sequences. IEEE Trans Robot 28(5): 1188-1197.
19. Simonyan, A Zisserman (2015) Very Deep Convolutional Networks for Large-Scale Image Recognition. Int Conf Learning Representations.
20. Patrick Wenzel, Rui Wang, Nan Yang, Qing Cheng, Qadeer Khan, et al. (2020) 4Seasons: A Cross-Season Dataset for Multi-Weather SLAM in Autonomous Driving. German Conf Pattern Recognition.
21. Y Cabon, N Murray, M Humenberger (2020) Virtual KITTI 2.
22. Gaidon Q, Wang Y Cabon, E Vig (2016) Virtual Worlds as Proxy for Multi-Object Tracking Analysis. IEEE Conf Comput Vision and Pattern Recognition pp. 4340-4349.
23. Santavas, I Kansizoglou, L Bampis, E Karakasis, A Gasteratos (2020) Attention! A Lightweight 2D Hand Pose Estimation Approach. IEEE Sensors Journal pp. 1-1.
24. Geiger P Lenz, C Stiller, R Urtasun (2013) Vision Meets Robotics: The KITTI Dataset. Int J Robot Res 32(11): 1231-1237.