5 Computer Vision Techniques for Urban Mobility

KOUROSH KHOSHELHAM

Abstract

This chapter provides an overview of computer vision techniques with applications in urban mobility and transport systems. Focusing on imagery and Light Detection and Ranging (LiDAR) point clouds as the main data modalities, the chapter reviews relevant computer vision tasks, including classification, segmentation, object detection and tracking. Example applications of these techniques to data captured by stationary sensors installed in the environment as well as mobile sensors onboard vehicles will then be discussed.

Keywords

Detection, tracking, classification, segmentation, localization, pose estimation

5.1 Introduction

The increasing prevalence of surveillance cameras in urban environments in recent years has provided an opportunity to develop new solutions to overcome challenges in urban mobility and transport systems. Cameras mounted on vehicles also offer the potential to sense the road environment and develop self driving capabilities which make urban mobility safer and more efficient. In addition to cameras, LiDAR sensors are becoming a preferred sensor for spatial perception in autonomous vehicles. These opportunities have led to a surge in the development of computer vision methods for automated interpretation of imagery and LiDAR point clouds with the ultimate aim of improving urban mobility.

This chapter reviews some promising applications of computer vision techniques for improving urban mobility. The focus will be on imagery and LiDAR point clouds as the more common data modalities for computer vision algorithms. Further, this chapter will focus on individual mobility, i.e. pedestrians, vehicles and cyclists. Other modes of mobility, such as freight, air, and maritime mobility, have received less attention from the computer vision research community, and are excluded from the present discussion.

https://doi.org/10.34727/2021/isbn.978-3-85448-045-7_5

While this chapter reviews example applications of computer vision techniques to urban mobility, it is not meant to serve as a classic review of the state of the art and is by no means exhaustive and comprehensive. Instead, the chapter aims to identify potential application areas where computer vision techniques can provide novel solutions to problems in urban mobility and transport systems. In the following, we first discuss common computer vision tasks for mobility applications, and then review promising examples of computer vision techniques applied to imagery and LiDAR point clouds captured by stationary sensors installed in the environment as well mobile sensors on board vehicles.

5.2 Common Computer Vision Tasks for Mobility Applications

Computer vision includes a wide range of algorithms developed to carry out specific tasks with the common goal of enabling a computer to understand the world by analyzing sensor observations in the form of images and point clouds. Common computer vision tasks for mobility applications include classification, segmentation, object detection, and tracking.

5.2.1 Classification

Classification is the task of assigning one or more category labels that identify the type of object or objects present in the data. The common approach to the classification of images and point clouds is supervised machine learning, where a mapping between the input data and the output category label is learned from a set of training examples. The category labels can be deterministic (hard labels) or probabilistic scores (soft labels).

Research on image classification made significant progress after the introduction of the ImageNet Challenge (Russakovsky et al., 2015) in 2010. The success of AlexNet (Krizhevsky et al., 2012) in the ImageNet 2012 Challenge led to the popularity of deep convolutional neural networks (CNNs) for image classification. Since then, many different CNN architectures have been proposed, such as VGG (Simonyan and Zisserman, 2015), GoogLeNet (Inception-v1) (Szegedy et al., 2015), and ResNet (He et al., 2016), which have achieved outstanding results on the ImageNet dataset. The classification of point clouds has achieved less success compared to image classification. State-of-the-art approaches to point cloud classification are either point-based methods, such as PointNet (Qi et al., 2017), or voxel-based methods, such as VoxNet (Maturana and Scherer, 2015).

5.2.2 Segmentation

Segmentation is the task of partitioning the data into segments that represent objects or parts thereof. A typical segmentation algorithm generates an output the same size as the input data, where each pixel or point is assigned a segment ID. If additionally, a category label is also assigned to each pixel or point, then the process is called semantic segmentation. The task of semantic segmentation is therefore a combination of segmentation and classification tasks.

Similar to classification, state-of-the-art segmentation and semantic segmentation methods are based on deep neural networks (Liu et al., 2019; Guo et al., 2020). The majority of these methods are based on supervised machine learning, where a deep network is trained using manually annotated images or point clouds available from public datasets.

5.2.3 Object Detection

Object detection is the task of localizing one or more objects of a certain category in the data. As such, object detection is a combination of classification and localization tasks. The localization is typically done by computing a bounding box around the object or a mask representing the object boundaries.

State-of-the-art approaches to object detection in imagery and point clouds are either based on region proposals or based on single shot classification and bounding box regression. Region proposal-based methods for object detection in images include Faster-RCNN (Ren et al., 2015) and Mask RCNN (He et al., 2017), and single shot methods include SSD (Liu et al., 2016) and the different versions of YOLO (Redmon et al., 2016). Methods for object detection in point clouds include PointRCNN (Shi et al., 2019), which is based on region proposals, and 3DSSD (Yang et al., 2020), which is a single shot method. These methods have been used for detecting vehicles, cyclists and pedestrians in images and LiDAR point clouds.

5.2.4 Tracking

In computer vision, tracking is the task of localizing an object in a sequence of data. Object tracking in a sequence of images or LiDAR scans usually involves detecting the object in the first image frame or LiDAR scan, and estimating its location in the subsequent frames or scans. The output of a tracking algorithm is the trajectory of the object in the sensor coordinate frame, which can be easily transformed to a trajectory on the ground by georeferencing the camera or the LiDAR sensor.

Recent methods for object tracking in images based on a convolutional Siamese network (Bertinetto et al., 2016; Wang et al., 2019) have achieved promising re-

sults in tracking people and vehicles. The Siamese network has also been extended for 3D tracking of pedestrians and cyclists in LiDAR data (Zarzar et al., 2019).

5.3 Computer Vision with Stationary Sensors

Recent advances in computer vision together with the prevalence of surveillance cameras installed in outdoor and indoor urban environments have made it possible to develop smart solutions for problems in mobility and urban transport. In the following pages, we review a few promising examples of such solutions made possible by computer vision methods. Most of the methods discussed in this section are based on imagery, as the use of stationary LiDAR sensors for monitoring urban environments is not currently common.

5.3.1 Pedestrian Detection and Tracking

Pedestrian detection and tracking using surveillance cameras and LiDAR sensors has been used in various urban mobility applications including pedestrian traffic management, prevention of overcrowding, origin-destination estimation, and monitoring intersections and pedestrian crossings. A practical application of pedestrian tracking in a video footage was shown by Kong et al. (2007) where the authors demonstrated that the tracking results can be used to proactively respond to incidents in a railway station. Another practical application of image based pedestrian tracking in indoor environments was demonstrated by Georgoudas et al. (2010) who developed an evacuation guidance system based on pedestrian tracking to prevent congestion during evacuations. For outdoor environments, image-based pedestrian tracking has been used to monitor intersections and provide useful information to improve the design of pedestrian crossings and adjust the signal timing (Malinovskiy et al., 2008).

A limitation of surveillance cameras for pedestrian tracking is their susceptibility to low light conditions especially in emergency situations in indoor environments. Li et al. (2019b) demonstrated the poor performance of color images for pedestrian origin-destination estimation during an emergency in a dark indoor environment, and proposed a deep convolutional network to fuse color, infrared and depth images for origin-destination estimation in emergency scenarios.

While pedestrian tracking using a single camera has been successfully applied in simple and small indoor and outdoor environments (Acharya et al., 2017), for large and more complex environments a multi-camera approach is preferred. Multi-camera pedestrian tracking includes the additional challenge of identity association across different camera views. Wu et al. (2020) formulated the 'identify' association as a graph-cut problem and showed an application of multi-camera

pedestrian tracking for analyzing the shopping behavior of customers in an indoor market hall.

Pedestrian detection and tracking in LiDAR data has also received a great deal of attention in recent years. Compared to cameras, LiDAR sensors are independent of ambient light and are less susceptible to poor lighting and adverse weather conditions. Zhao et al. (2018) demonstrated the application of pedestrian tracking using a roadside LiDAR sensor to infer the crossing intention of pedestrians.

Current methods for pedestrian detection and tracking in images and LiDAR data are successful in less crowded scenes where individual pedestrians are clearly visible. For crowded scenes, where extracting the complete trajectories of individual pedestrians may not be feasible, extracting global parameters such as crowd density, global velocity (Yi et al., 2015), and congestion is more convenient.

5.3.2 Crowd Congestion Classification

Crowd congestion information automatically extracted from surveillance images in real time provides valuable insights for the management of busy transport hubs especially during peak commute times. Crowd congestion is usually measured as the average occupancy area available per person, commonly referred to as level of service. As such, it can be estimated by detecting and counting the pedestrians in surveillance images and computing the crowd density (Ryan et al., 2015). However, in crowded scenes, where pedestrians are partly occluded in the images, counting, and density estimation will be inaccurate and may lead to incorrect congestion classification results.

An alternative approach is to directly classify local image regions into different congestion classes based on crowd appearance features. Li et al. (2019a) trained a long short term memory (LSTM) network using manually labelled images of different crowd densities to classify image patches corresponding to a grid on the ground and generate a level of service map of a railway platform (Figure 5.1). The resulting map overlaid on a 3D model of the platform provides an effective visualization of both spatial and temporal variations of congestion classes in real time. To avoid the influence of occlusion in a single view, Li et al. (2020) extended this approach to multiple views by classifying image patches corresponding to a grid in each camera view and combining the classification results using an ensemble combination rule. This multi-view approach was shown to produce a more accurate level of service map than that obtained from each individual view.

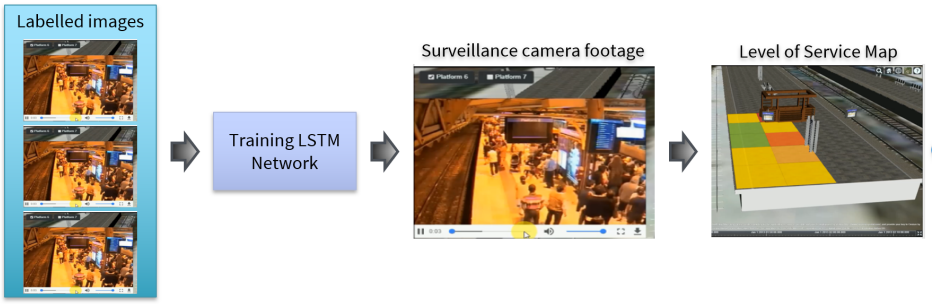

Figure 5.1: Direct generation of crowd congestion map from surveillance images.

5.3.3 Parking Occupancy Detection

Parking occupancy detection using surveillance cameras provides a low-cost yet accurate and reliable solution for smart parking systems in crowded cities. The common approach is to train a binary classifier to classify image regions corresponding to parking spaces as either occupied or vacant. Early works such as True (2007) used hand crafted features based on the appearance of vehicles and achieved modest accuracies. But recent advances in feature learning using deep convolutional networks made it possible to achieve much higher accuracies in vehicle detection and determining the occupancy of parking spaces. Valipour et al. (2016) trained a deep VGG network using labelled images from PKLot dataset (De Almeida et al., 2015) and reported an occupancy detection accuracy of 99 % on a test set from the same dataset. Acharya et al. (2018) investigated the feasibility of transfer learning, where a deep network trained on a public dataset such as PKLot is applied to images captured in a different parking setting. They tested this approach using an SVM classifier plugged into a VGG network and reported an accuracy of 97 % for detecting the occupancy of parking spaces. Chapter 11 provides a tutorial on the transfer learning approach to image-based parking occupancy detection using a ResNet architecture.

5.3.4 Detection of Anomalous Driving Behaviors

An interesting application of computer vision techniques in urban mobility is automated detection of anomalous driving behaviors, such as swerving, speeding, and crossing solid lines, in surveillance images. While methods for detecting different anomalous behaviors may be different, a common ingredient of these methods is vehicle detection, tracking and reconstruction of vehicle trajectories. An early work on image-based analysis of driving behaviors is the work of Song et al. (2014), who used simple background elimination and feature point extraction to detect and track vehicles in video footages of several roads in Xian,

Shanghai, and Fuzhou. They used the reconstructed vehicles trajectories to estimate the speed and identify various anomalous behaviors such as lane changing, sudden stopping, and sudden slowing down. Zheng et al. (2019) proposed a taxonomy of anomalous driving behaviors and developed a vehicle detection and tracking system based on Mask RCNN (He et al., 2017) to detect speed anomalies, solid line crossing, and vehicles entering restricted zones such as a bus lane. They also proposed a web mapping application to visualize anomalous driving behaviors on different roads as a guide for vulnerable road users such as cyclists and pedestrians.

5.4 Computer Vision with Mobile Sensors

The widespread interest in autonomous vehicles in recent years has resulted in the development of computer vision techniques for spatial perception of road environments using cameras and LiDAR sensors on board vehicles. This section reviews a few examples of promising applications of computer vision techniques applied to imagery and point clouds captured by vehicle-borne cameras and LiDAR sensors.

5.4.1 Driving Scene Perception

Automated perception and understanding of the driving scene is a critical capability for the successful operation of fully autonomous vehicles. A first computer vision task for autonomous vehicles is to detect the road boundaries and lane markings. Many modern vehicles already have the lane detection and lane keeping capability on well marked roads. The challenge, however, is the detection of road and lane boundaries on unmarked and weakly marked roads. The KITTI Road Detection Benchmark (Fritsch et al., 2013) provides an evaluation and comparison of road detection methods based on images and LiDAR data on several challenging datasets. When road markings are not clearly visible in the data, the fusion of images and LiDAR point clouds can provide more reliable detection results. Chen et al. (2019) train a convolutional network to learn and fuse image and LiDAR features to detect road boundaries, and achieve state of the art performance on KITTI road detection dataset. Prior knowledge and existing maps can also be used to support road and lane detection in sensor data. Wang et al. (2020) take advantage of road information from OpenStreetMaps and combine it with image features in a search-based optimization algorithm to estimate the correct location of lane boundaries.

Another important computer vision task for autonomous vehicles is the recognition of traffic signs. State of the art deep learning methods for image classification generally achieve high accuracies in traffic sign recognition in images. The German traffic sign recognition benchmark (Stallkamp et al., 2012) demon-

strated that deep convolutional networks can achieve correct classification rates up to 99.46 % on test images of various traffic signs.

Detection of vehicles, pedestrians and cyclists in the road environment is another important computer vision task for autonomous vehicles. It is a particularly challenging task due to the dynamic nature of objects which can result in occlusion and obscure images. The KITTI Vision Benchmark Suite (Geiger et al., 2012) provides a dataset comprising imagery and LiDAR data of vehicles, pedestrians, and cyclists at three levels of occlusion: fully visible (easy), partly occluded (moderate), and difficult to see (hard). The results of the benchmark show that current methods are generally better at 2D detection than 3D detection. For example, the current top performing method for 2D detection of pedestrians achieves an average precision of 90.50 %, 83.06 %, and 78.35 % on easy, moderate, and hard test samples respectively, whereas the best average precision for 3D pedestrian detection is only 53.10 %, 45.37 %, and 41.47 % for easy, moderate, and hard test samples. Also, the detection of vehicles seems to be an easier task, while the detection of pedestrians and cyclists is a greater challenge. For example, the current best average precision for 3D car detection in KITTI Benchmark is 82.33 % on moderate test samples, whereas for 3D detection of cyclists and pedestrians the best average precision on moderate samples drops to 71.86 % and 45.35 %, respectively.

Other computer vision tasks related to autonomous driving include the detection of road incidents and road surface conditions. Levering et al. (2020) proposed a taxonomy of unsigned road incidents and developed a deep learning model to recognize eight types of road incidents in driver view images, namely vehicle crash, tree-fall, fire, landslide, collapse, flood, snow, and animal on road. Pena-Caballero et al. (2020) proposed a system to detect potentially hazardous road surface conditions such as potholes and cracks using driver view images. These methods can be used in a crowd-sourcing approach to collect information about road conditions and use centralized or decentralized communication systems to disseminate the information among all road users.

5.4.2 Generation of High-definition Maps of Road Environments

High-definition (HD) maps are highly detailed 3D maps containing the 3D location of all traffic signs, traffic lights, trees, and every relevant object in the road environment. HD maps are considered an essential component of fully autonomous vehicles. An HD map enables the autonomous vehicle to localize itself accurately with respect to the road environment and recognize and react to events on the road, which might not be detected by the sensors on board the vehicle.

While there is currently no standard specifying the format and structure of HD maps, it is widely accepted that the raw material for the generation of HD maps are 3D data, such as LiDAR point clouds, with semantic information representing

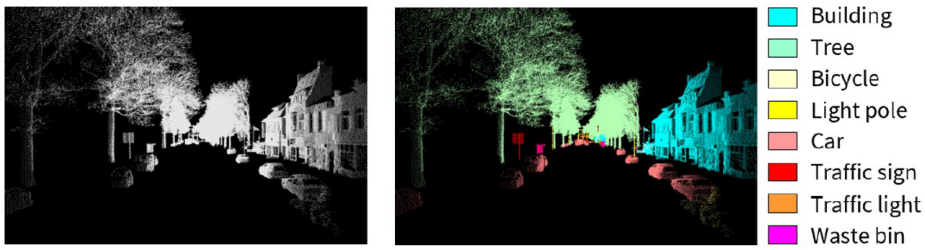

Figure 5.2: An example of raw point cloud collected by a mobile LiDAR sensor (left), and the classified point cloud (right).

the type of objects present in the data. Efficient generation of HD maps requires automated recognition and classification of various objects in the point cloud. Figure 5.2 shows an example of raw 3D point cloud acquired by a mobile LiDAR sensor and the classified point cloud containing semantic information about the type of objects present in the environment.

Classification methods applied to point clouds of road environments have thus far been less successful due to the complexity of the objects involved. For example, the current top performing approach in the Paris-Lille-3D benchmark (Roynard et al., 2018) achieves a mean intersection over union (IoU) score of only 82.7 % (Boulch et al., 2020). The most complex objects for classification are small objects with intra-class variability such as traffic signs, traffic lights, and light poles. For instance, the highest mean IoU for the recognition of poles in the the Paris-Lille-3D benchmark at present is 79.7 % (Luo et al., 2020). A fundamental problem contributing to the poor performance of classification methods applied to point clouds, when compared to images, is the scarcity of labelled data for training. The Paris-Lille-3D dataset, which is one of the largest urban point cloud datasets, contains 2479 labelled segments across 50 categories (Roynard et al., 2018), that is an average 50 training samples per category. Other LiDAR datasets for autonomous driving, such as KITTI (Geiger et al., 2012), nuScenes (Caesar et al., 2019), and Waymo Open (Sun et al., 2020), have annotations for fewer object categories. In comparison, the ImageNet Challenge dataset contains roughly 1000 training images in each of 1000 categories (Krizhevsky et al., 2012). The limited availability of training samples from urban point clouds is mainly due to the complexity of annotating point clouds as compared to image labelling.

5.4.3 Vehicle Localization

Estimating the location of the vehicle with respect to a map is a basic requirement for autonomous navigation. While the Global Navigation Satellite System

(GNSS) is the primary technology for vehicle localization, in urban environments where GNSS signals are not available, e.g. urban canyons and tunnels, computer vision techniques using images and LiDAR data can be used to estimate the location of the vehicle. Vehicle localization methods based on imagery and LiDAR point clouds can be divided into two categories: local motion estimation and global position estimation (Khoshelham and Ramezani, 2017).

In local motion estimation, the position of the vehicle is determined by estimating its motion with respect to a previously known position. Local motion estimation methods using imagery and LiDAR data are mainly based on visual odometry (Ramezani and Khoshelham, 2018; Ramezani et al., 2018) and simultaneous localization and mapping (SLAM) (Bresson et al., 2017). A major limitation of visual odometry and SLAM approaches to vehicle localization is the drift of the estimated trajectory caused by the accumulation of errors in each local motion estimation step. Overlap detection and loop closing methods, such as OverlapNet (Chen et al., 2020), can be used to correct the drift. However, for vehicle localization correct location estimates are needed in real time and correction of the trajectory with some delay is not practical.

In global position estimation, the position of the vehicle is estimated directly in a global reference coordinate frame by matching the images or LiDAR scans with a georeferenced source of spatial data. Image-based pose regression methods, such as PoseNet (Kendall et al., 2015), estimate the pose of the camera by learning a regressor from a set of images with known pose. LiDAR-based methods, such as L3Net (Lu et al., 2019), learn correspondences between a current LiDAR scan and a set of pre-existing LiDAR scans of the environment to estimate the position of the vehicle. Other methods detect landmarks, such as road signs (Ghallabi et al., 2019) and curbs (Wang et al., 2017), in LiDAR data and match these with a pre-existing map to estimate the location of the vehicle. The prerequisite for all these approaches is the availability of a set of georeferenced images, LiDAR scans, or HD maps of the environment.

Computer vision approaches to vehicle localization are generally considered complementary to GNSS rather than competitive. As such, location estimates from imagery and LiDAR data are often fused with GNSS measurements when available. Gao et al. (2015) and Ilci and Toth (2020) propose methods for the integration of LiDAR localization methods with GNSS and inertial measurements.

5.5 Concluding Remarks

The potential of computer vision techniques for urban mobility applications has been demonstrated in many recent works as reviewed in this chapter. However, a few challenges still remain to be addressed. The first challenge is the practicality of machine learning approaches in real scenarios. Most of the existing methods

are based on supervised learning, which requires an off-line training phase and adequate training examples. But, in many practical applications where a plug and play solution is needed unsupervised or semi supervised learning models are preferred. Transfer learning using pre-trained deep networks is a promising solution for image-based methods. However, at present pre-trained models for LiDAR data are scarce and have poor transferability. Generative models and training by synthetic samples are potential approaches to unsupervised and semi-supervised learning which are worth further exploration.

A related challenge for the application of computer vision to urban mobility is the geographical diversity and scene adaptation. Most existing methods are scene-dependent. For example, a machine learning model trained on a dataset captured in Paris might perform poorly on data captured in Melbourne. Domain adaptation methods such as sample weighting and distribution alignment, e.g. using adversarial training, have received little attention so far and are worth further investigation.

Robustness to poor lighting and adverse weather conditions is another important challenge for computer vision methods. Recent research is paying more attention to the development of more robust computer vision methods. This is further promoted by the development of public datasets for autonomous driving which provide annotated images and LiDAR data captured in rain, snow, and night time; see e.g. nuScenes (Caesar et al., 2019) and Canadian Adverse Driving Conditions (CADC) dataset (Pitropov et al., 2020).

Bibliography

Acharya, D., Khoshelham, K., and Winter, S. (2017). Real-time detection and tracking of pedestrians in CCTV images using a deep convolutional neural network. In *Proceedings of Research@Locate17, 3-6 April 2017, Sydney, Australia*, volume 1913, pages 31–36.

Acharya, D., Yan, W., and Khoshelham, K. (2018). Real-time image-based parking occupancy detection using deep learning. In *Proceedings of Research@Locate18, 9-11 April 2018, Adelaide, Australia*, volume 2087, pages 33–40.

Bertinetto, L., Valmadre, J., Henriques, J. F., Vedaldi, A., and Torr, P. H. (2016). Fully-convolutional siamese networks for object tracking. In *European Conference on Computer Vision*, pages 850–865. Springer.

Boulch, A., Puy, G., and Marlet, R. (2020). Fkaconv: Feature-kernel alignment for point cloud convolution. *arXiv preprint arXiv:2004.04462*.

Bresson, G., Alsayed, Z., Yu, L., and Glaser, S. (2017). Simultaneous localiza-

tion and mapping: A survey of current trends in autonomous driving. *IEEE Transactions on Intelligent Vehicles*, 2(3):194–220.

Caesar, H., Bankiti, V., Lang, A. H., Vora, S., Liong, V. E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., and Beijbom, O. (2019). nuscenes: A multimodal dataset for autonomous driving. *arXiv preprint arXiv:1903.11027*.

Chen, X., Läbe, T., Milioto, A., Röhling, T., Vysotska, O., Haag, A., Behley, J., Stachniss, C., and Fraunhofer, F. (2020). Overlapnet: Loop closing for lidar-based slam. In *Proc. of Robotics: Science and Systems (RSS)*.

Chen, Z., Zhang, J., and Tao, D. (2019). Progressive lidar adaptation for road detection. *IEEE/CAA Journal of Automatica Sinica*, 6(3):693–702.

De Almeida, P. R., Oliveira, L. S., Britto Jr, A. S., Silva Jr, E. J., and Koerich, A. L. (2015). Pklot–a robust dataset for parking lot classification. *Expert Systems with Applications*, 42(11):4937–4949.

Fritsch, J., Kuehnl, T., and Geiger, A. (2013). A new performance measure and evaluation benchmark for road detection algorithms. In *16th International IEEE Conference on Intelligent Transportation Systems (ITSC 2013)*, pages 1693–1700. IEEE.

Gao, Y., Liu, S., Atia, M. M., and Noureldin, A. (2015). Ins/gps/lidar integrated navigation system for urban and indoor environments using hybrid scan matching algorithm. *Sensors*, 15(9):23286–23302.

Geiger, A., Lenz, P., and Urtasun, R. (2012). Are we ready for autonomous driving? The KITTI vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Georgoudas, I. G., Sirakoulis, G. C., and Andreadis, I. T. (2010). An anticipative crowd management system preventing clogging in exits during pedestrian evacuation processes. *IEEE Systems Journal*, 5(1):129–141.

Ghallabi, F., El-Haj-Shhade, G., Mittet, M.-A., and Nashashibi, F. (2019). Lidar-based road signs detection for vehicle localization in an hd map. In *2019 IEEE Intelligent Vehicles Symposium (IV)*, pages 1484–1490. IEEE.

Guo, Y., Wang, H., Hu, Q., Liu, H., Liu, L., and Bennamoun, M. (2020). Deep learning for 3d point clouds: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

He, K., Gkioxari, G., Dollár, P., and Girshick, R. (2017). Mask R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2961–2969.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Ilci, V. and Toth, C. (2020). High definition 3d map creation using gnss/imu/lidar sensor integration to support autonomous vehicle navigation. *Sensors*, 20(3):899.

Kendall, A., Grimes, M., and Cipolla, R. (2015). Posenet: A convolutional network for real-time 6-dof camera relocalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2938–2946.

Khoshelham, K. and Ramezani, M. (2017). Vehicle positioning in the absence of gnss signals: Potential of visual-inertial odometry. In *Joint Urban Remote Sensing Event*. IEEE.

Kong, S., Sanderson, C., and Lovell, B. C. (2007). Classifying and tracking multiple persons for proactive surveillance of mass transport systems. In *2007 IEEE Conference on Advanced Video and Signal Based Surveillance, AVSS 2007*, pages 159–163, Los Alamitos, CA, USA. IEEE Computer Society.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105.

Levering, A., Tomko, M., Tuia, D., and Khoshelham, K. (2020). Detecting unsigned physical road incidents from driver-view images. *IEEE Transactions on Intelligent Vehicles*.

Li, Y., Khoshelham, K., Sarvi, M., and Haghani, M. (2019a). Direct generation of level of service maps from images using convolutional and long short-term memory networks. *Journal of Intelligent Transportation Systems*, 23(3):300–308.

Li, Y., Sarvi, M., and Khoshelham, K. (2019b). Pedestrian origin-destination estimation in emergency scenarios. In *9th International Conference on Fire Science and Fire Protection Engineering (ICFSFPE)*, pages 1–5. IEEE.

Li, Y., Sarvi, M., Khoshelham, K., and Haghani, M. (2020). Multi-view crowd congestion monitoring system based on an ensemble of convolutional neural network classifiers. *Journal of Intelligent Transportation Systems*, pages 1–12.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., and Berg, A. C. (2016). Ssd: Single shot multibox detector. In *European Conference on Computer Vision*, pages 21–37. Springer.

Liu, X., Deng, Z., and Yang, Y. (2019). Recent progress in semantic image segmentation. *Artificial Intelligence Review*, 52(2):1089–1106.

Lu, W., Zhou, Y., Wan, G., Hou, S., and Song, S. (2019). L3-net: Towards learning based lidar localization for autonomous driving. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6389–6398.

Luo, H., Chen, C., Fang, L., Khoshelham, K., and Shen, G. (2020). Ms-rrfsegnet: Multiscale regional relation feature segmentation network for semantic segmentation of urban scene point clouds. *IEEE Transactions on Geoscience and Remote Sensing*.

Malinovskiy, Y., Wu, Y.-J., and Wang, Y. (2008). Video-based monitoring of pedestrian movements at signalized intersections. *Transportation Research Record*, 2073(1):11–17.

Maturana, D. and Scherer, S. (2015). Voxnet: A 3d convolutional neural network for real-time object recognition. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 922–928. IEEE.

Pena-Caballero, C., Kim, D., Gonzalez, A., Castellanos, O., Cantu, A., and Ho, J. (2020). Real-time road hazard information system. *Infrastructures*, 5(9):75.

Pitropov, M., Garcia, D., Rebello, J., Smart, M., Wang, C., Czarnecki, K., and Waslander, S. (2020). Canadian adverse driving conditions dataset. *arXiv preprint arXiv:2001.10117*.

Qi, C. R., Su, H., Mo, K., and Guibas, L. J. (2017). Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 652–660.

Ramezani, M. and Khoshelham, K. (2018). Vehicle positioning in gnss-deprived urban areas by stereo visual-inertial odometry. *IEEE Transactions on Intelligent Vehicles*, 3(2):208–217.

Ramezani, M., Khoshelham, K., and Fraser, C. (2018). Pose estimation by omnidirectional visual-inertial odometry. *Robotics and Autonomous Systems*, 105:26–37.

Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 779–788.

Ren, S., He, K., Girshick, R., and Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99.

Roynard, X., Deschaud, J.-E., and Goulette, F. (2018). Paris-lille-3d: A large and high-quality ground-truth urban point cloud dataset for automatic segmentation and classification. *The International Journal of Robotics Research*, 37(6):545–557.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. (2015). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252.

Ryan, D., Denman, S., Sridharan, S., and Fookes, C. (2015). An evaluation of crowd counting methods, features and regression models. *Computer Vision and Image Understanding*, 130:1–17.

Shi, S., Wang, X., and Li, H. (2019). PointRCNN: 3d object proposal generation and detection from point cloud. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–779.

Simonyan, K. and Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*.

Song, H.-S., Lu, S.-N., Ma, X., Yang, Y., Liu, X.-Q., and Zhang, P. (2014). Vehicle behavior analysis using target motion trajectories. *IEEE Transactions on Vehicular Technology*, 63(8):3580–3591.

Stallkamp, J., Schlipsing, M., Salmen, J., and Igel, C. (2012). Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition. *Neural Networks*, 32:323–332.

Sun, P., Kretzschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., Guo, J., Zhou, Y., Chai, Y., Caine, B., et al. (2020). Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2446–2454.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9.

True, N. (2007). Vacant parking space detection in static images. *University of California, San Diego*, 17:659–662.

Valipour, S., Siam, M., Stroulia, E., and Jagersand, M. (2016). Parking-stall vacancy indicator system, based on deep convolutional neural networks. In *2016 IEEE 3rd World Forum on Internet of Things (WF-IoT)*, pages 655–660. IEEE.

Wang, L., Zhang, Y., and Wang, J. (2017). Map-based localization method for autonomous vehicles using 3d-lidar. *IFAC-PapersOnLine*, 50(1):276–281.

Wang, Q., Zhang, L., Bertinetto, L., Hu, W., and Torr, P. H. (2019). Fast online object tracking and segmentation: A unifying approach. In *Proceedings of the*

IEEE Conference on Computer Vision and Pattern Recognition, pages 1328–1338.

Wang, X., Qian, Y., Wang, C., and Yang, M. (2020). Map-enhanced ego-lane detection in the missing feature scenarios. *arXiv preprint arXiv:2004.01101*.

Wu, X., Winter, S., and Khoshelham, K. (2020). Multi-camera tracker for monitoring pedestrians in enclosed environments. In *International Conference on Tools with Artificial Intelligence, ICTAI 2020*.

Yang, Z., Sun, Y., Liu, S., and Jia, J. (2020). 3DSSD: Point-based 3d single stage object detector. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11040–11048.

Yi, S., Li, H., and Wang, X. (2015). Pedestrian travel time estimation in crowded scenes. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3137–3145.

Zarzar, J., Giancola, S., and Ghanem, B. (2019). Efficient tracking proposals using 2d-3d siamese networks on lidar. *arXiv preprint arXiv:1903.10168*.

Zhao, J., Xu, H., Wu, J., Zheng, Y., and Liu, H. (2018). Trajectory tracking and prediction of pedestrian's crossing intention using roadside lidar. *IET Intelligent Transport Systems*, 13(5):789–795.

Zheng, X., Wu, F., Chen, W., Naghizade, E., and Khoshelham, K. (2019). Show me a safer way: Detecting anomalous driving behavior using online traffic footage. *Infrastructures*, 4(2):22.