Development in Vision-Based On-Road Behaviors Understanding

Subjects: Imaging Science & Photographic Technology | Computer Science, Artificial Intelligence
Contributor: Redouane Khemmar, Rim Trabelsi, Benoit Decoux, Remi Boutteau

On-road behavior analysis is a crucial and challenging problem in the autonomous driving vision-based area. Several endeavors have been proposed to deal with different related tasks and it has gained wide attention recently. Much of the excitement about on-road behavior understanding has been the labor of advancement witnessed in the fields of computer vision, machine, and deep learning. Remarkable achievements have been made in the Road Behavior Understanding area over the last years.

Keywords: holistic on-road behavior analysis ; driver-road interaction ; driving status analysis ; road scene understanding

1. Introduction

Vision-based Advanced Driver-Assistance Systems (ADAS) and Autonomous Vehicles (AV) development has been witnessing a slow yet impressive growth over the past two decades for most of its modules including sensing hardware [1], perception [2], mapping and path planning [3], and scene understanding [4]. Despite the large amount of research and development endeavors, the area still lacks comprehensive approaches and/or systems which provide high-level road scene analysis that includes holistic description of the AV'surrounding. By contrast to most of the ADAS and AV vision-based tasks, there are scarce resources for broad road scene understanding, which is mainly due to the paucity of benchmarks, standards, theoretical fundamentals, and expressly research papers that urge the community to tackle this task.

2. Importance and Challenges of On-Road Behaviors Analysis

Road crashes kill nearly "1.25 million people each year, on average 3287 deaths a day and around 50 million are injured or disabled", reports the Association for Safe International Road Travel (ASIRT). By 2030, fatal traffic crashes are expected to become the fifth leading cause of death worldwide. ITS (Intelligent Transportation Systems) is investigating to find smarter solutions that allow bolder ADAS and AV systems in order to reduce road fatalities.

To do so, many vision-based components have been developed and significant advances have been made on semantic segmentation [5], object detection [6], mapping and localization [3], etc. Yet, such tasks have not really addressed the challenges in high-level understanding and correspond only to the first step of understanding, thus the importance of on-road scene analysis. In fact, detecting/localizing participants in a given scene and parsing them to the semantic classes is just a mid-level analysis step.

Despite the ITS community always asking "what are the difficulties and challenges in on-road behaviors analysis?", this question has not been raised seriously so far in the literature or even been generalized. As vision-based understanding related tasks have different goals and constraints, integrating a set of these tasks under the same framework implies imperatively a higher complexity. Aside from common issues in image and video analysis tasks such as sudden and gradual illumination changes and viewpoints variation, the specificity of road environments involves more aspects including temporal and spatial patterns intra-class variations, cluttered scenes, and the curse of data labeling for high-level analysis.

3. RBA Components and Milestones over the Last 20 Years

3.1. Situational Awareness

Situational awareness is a fundamental ingredient for successful on-road scene analysis. In fact, while dealing with the perception of road environments, including objects, participants, behaviors, and activities, people need to take into

consideration their meaning, and their future status in conjunction with space and time.

For example, Kooij et al. propose in [7] a pedestrian situational awareness aiming to predict a pedestrian's path in the AV domain. To anticipate the decision of pedestrians, they assume that it is influenced by three factors: (i) the presence of an approaching vehicle on a collision course, (ii) the pedestrian's safety awareness, and (iii) the specificity of the environment. To incorporate these factors, latent states have been built on switching linear dynamical systems and dynamic Bayesian networks to anticipate changes in pedestrian paths. The situational awareness evaluates whether the pedestrian at a given instant t has seen the vehicle previously (before t)) through the estimation of the distance between vehicle and pedestrian at the expected point of closest approach head orientation and his distance to the road curbside. The ultimate goal is to predict the intention of a pedestrian to laterally cross the road, which is highly related to most of the pedestrian fatalities inroads according to accident analysis reported in [8]. For a similar objective, in [9] the crossing behavior, i.e., the pedestrians' intention to cross, is predicted. Researchers consider here two classification tasks of static environmental context and passenger action. Due to the lack of annotation of the JAAD dataset proposed, they make use of weakly supervised learning with CNN (Convolutional Neural Network) features through an AlexNet architecture to identify visual attributes for both tasks [10]. The considered classes for the environmental contextual recognition task include the following elements: narrow, wide, pedestrian crossing sign, zebra crossing, stop sign, traffic light, and parking lot. As for the pedestrian action classification, only looking and crossing classes have been considered. Their experimental results prove that using only the pedestrian action information can predict 40% of the observed crossing behavior, yet, adding situational awareness information/context improves the results by 20%.

3.2. Driver-Road Interaction

Understanding the interactions occurring between drivers and road users is crucial toward a holistic RBA. Recently, the ITS community has brought attention to this issue by developing new vision-based models that capture the interaction between road users and the ego vehicle.

Owing to the newly published HDD dataset [11], Li et al. propose in [12] a spatio-temporal 3D-aware framework by means of GCN (Graph Convolution Networks). For this purpose, participants and objects of the road scene are categorized into two sets: Thing objects, such as pedestrians and cars, and Stuff objects, such as traffic lights and road surface marking. As an example, interactions with Thing objects may include stopping to let pedestrians cross and deviation for parked vehicles, and examples for interactions with Stuff objects may include changing or merging a lane. To represent the interaction between these sets of objects and the ego vehicle, two kinds of graphs have been proposed: Ego-Thing Graph and Ego-Stuff Graph to represent the ego vehicle interaction with Thing and Stuff objects, respectively by extending the Ego-Thing Graph proposed in [13].

Previously, in 2000, using traditional machine learning tools in Ref.[14], HMM (Hidden Markov Model) was leveraged to recognize the tactical driver behavior. To model the states of the ego vehicle and all the elements and participants in the road environment, a single node was used for each one and then all the nodes were encoded into a state vector. This approach is not reporting state-of-the-art results so far, but it still inspires recent endeavors like those stated earlier.

3.3. Road Scene Understanding

ITS and computer vision communities have exerted extensive effort to analyze the road scene [15][16][17]. As hinted to earlier, traditional approaches were handcrafted or machine learning-based and they were designed to deal with an elementary task such as detection recognition or segmentation for a specific kind of object lane [18], traffic lights [19], and pedestrians [20]. Before 2013, it was not obvious how to get a holistic understanding of road environments. With the advent of the revolutionary deep neural networks, it was made possible to get a comprehensive understanding of the road scene. In fact, deep learning provides more automatic models, which allow combining many tasks in a single framework.

For example, Ref. [21] proposes an encoder-decoder architecture that provides an end-to-end road scene understanding. The encoder is a CNN-based network similar to VGGNet [10]. Similarly, the decoder is a CNN network with two streams that upsamples the features and then fuses the information of both streams to the decoder network, which outputs the score of each pixel with regards to the predefined classes. By reporting qualitative and quantitative evaluation done only over the CamVid dataset [22], this encoder-decoder network is able to provide higher results for semantic segmentation and positioning in terms of CA (Class Accuracy) at MIoU (Mean Intersection over Union).

Under the same context, Ref. [5] deals with a more challenging task: recovery of occluded vehicles, in addition to the amodal segmentation task. A novel multi-task approach was introduced to incorporate the two tasks under the same framework. In another word, two networks were introduced, called (i) segmentation completion and (ii) appearance

recovery. The first one aims to produce the recovered segmentation mask. Afterward, the obtained mask is used for the second network to produce the occluded parts of vehicles. As a result, the invisible regions are restored/painted back to the foreground of the original image. Experiments were done with the Occluded Vehicle Dataset (OVD) and metrics used to evaluate the recovered segmentation mask including recall, precision, F1 score, IuO, the per-pixel L1 error and the per-pixel L2 error were used. As for the recovered appearance of the vehicle task, the L1 and L2 errors have been used along with two new metrics: the ICP (Inception Conditional Probability) and the SS (Segmentation Score). Overall, the obtained results demonstrate similar performance to Deeplab [23] without the iterative refinement for the segmentation with a better capability of generating the invisible parts.

As for the work proposed in Ref. [24], researchers focus on urban scene analysis and road detection and evaluate their proposal over KITTI [25] and LabelMeFacade [26] datasets. The major contribution is the presentation of new convolutional patch networks learned for pixel-wise labeling in order to classify image patches. More details regarding the proposed CNN can be found in Section 5.

3.4. Trajectories Forecast

Trajectories forecast in road environments consists of predicting the path that a moving vehicle, pedestrian, or any other road agent follows through road space as a function of time. The extracted trajectories from a visual data are less mastered with relation to ADAS and AV contexts. It is highlighted its importance and the methods developed by the computer vision community that can be useful in such circumstances. In addition to the prediction of future paths of road participant, this RBA component is of crucial importance and might be used to supply other components by modeling traffic representation in real-time. Thus, this task is useful for performing safe ADAS and AV systems not only by improving all the vision-based components but also by using it in the management of congestion and vehicle routing.

For instance, Ref. [27] proposes an approach named DROGON for “Deep RObust Goal-Oriented trajectory prediction Network” that forecasts future trajectories of vehicles by taking into consideration their behavioral intention, which is commonly called causal reasoning. A conditional prediction model has been built to solve this issue. Three steps have been followed to train such a model. The first step is to infer the interaction of vehicles with other road agents. Then, the second and the third steps consist in estimating their intention and thus by computing the intention’s probability distribution based on the inferred interaction and the causal reasoning, respectively. To evaluate their approach, Choi et al. also introduce a new dataset acquired at four-way intersections. Through qualitative and quantitative evaluation using ADE (Average Distance Error) and FDE (Final Distance Error) metrics, the DROGON framework has demonstrated an efficient performance for predicting and generating trajectories over the state-of-the-art.

A near-term trajectories forecast approach is also proposed by Chandra et al. in [28]. They tackle more challenging contexts, mainly heterogeneous traffic where road agents are very varied for, e.g., pedestrians, cars, bicycles, trucks, buses, motorcycles, etc. The major contribution is the representation of heterogeneous interactions between different kind of agents. An RNN-CNN framework called TraPHic is introduced in order to automatically predict the trajectories and does so by taking into consideration the challenges of heterogeneous contexts related to different textures, shapes, motions, and activities of each road participants.

3.5. Driving Activities and Status Analysis

Driver’s status and activities analysis are crucial because one of the leading cause of traffic accidents is related to driver’s inattention, aggressive maneuvering, or drowsiness. Detecting or predicting strange maneuvers or status may help avoiding fatal crashes by keeping a distance to identified aggressive drivers.

In the literature, Ref. [29] proposes a novel framework named GraphRQI to learn, in a supervised fashion, third party driver behavior. Based on the approach proposed in Ref. [28] (detailed in the previous section), road participants trajectories are used to pinpoint the general traits of the driver, i.e., conservative or aggressive, which inherently affects the trajectories of the surroundings agents. By means of GCNs trained on the TRAF [28] and Argoverse [30] datasets, driving status has been classified into six classes: reckless, careful, timid, impatient, threatening, and cautious.

In a similar way, Ref. [31] classifies the ego vehicle behavior through scene understanding-based CNN and robust temporal representation-based LSTM (Long Short Term Memory). The key element is the incorporation of scene context features related to weather and road characteristic, places, types, and surface conditions. Experimental results shown on the HDD dataset and a newly collected dataset prove that the proposed scene classification boosts the understanding of driver behavior.

Multi-modality cues have been extensively used with regard to driver maneuvers classification for example, by taking into consideration only the temporal aspect, Ref. [32] presented a fancy architecture based on Gated Recurrent Fusion Unit to model multi-modal data coming from two streams: video and CAN bus. Novel gating functions have been introduced in order to represent the exposure of each data stream at each frame in order to provide an adaptive data fusion. To recognize the action that occurred during the driving scenarios over the HDD dataset, Ref. [33] employs a CNN-based architecture with a triplet loss as a space embedding regularizer to limit the problem of overfitting encountered with HDD pre-trained models. Extensive evaluation done on the HDD dataset proves that softmax plus triplet loss achieves a better performance on minority classes, proving a better generalization of the trained network.

Following a multi-modal framework as well, Ref. [34] presents DBUS (Driving Behavior Understanding System), which makes use of image sequences synchronized with GPS/IMU signals. The perks of DBUS is the recognition of the driver's attention and intention along with the maneuver. A unified model has been proposed to jointly analyze the attention and the intention of the driver activity. Even though the proposal is original, its efficiency was not evaluated over state-of-the-art benchmarks and it has only been tested on a non publicly available dataset.

Using the same concept of multi-modal and multi-level understanding, Ref. [11] recognizes driver behaviors through different layers including attention, cause, stimulus-driven action, and goal-oriented action. Contrary to Ref. [34], the proposed approaches based on standard CNN and LSTM have been evaluated over the HDD benchmarks devoted to learning the driver's activities and causal reasoning. The same model has been reproduced in [35] where Chen et al. propose to classify driving behavior through new modes, color, and depth sequences coming from video and LiDAR sensors, respectively.

3.6. Holistic Understanding

Holistic understanding aims at jointly dealing with different complementary tasks within the same framework for RBA. In the literature, there was a lack of serious endeavors until 2015. Jain et al. proposed the first-ever framework that incorporates several levels of understanding to predict maneuvers of the ego vehicle [36]. Both inside and outside contexts (with regards to the vehicle) have been taken into consideration. Researchers consider the visual representation for the driver's facial expression and motion. As for the outside context, scene perception has been investigated to anticipate a driver's maneuver through multi-modal features such as image sequences, GPS, road maps, speed, and events. Auto-regressive HMM models have been proposed to generate the latent states of the driver [37]. Owing to the uniqueness of the approach, promising results have been obtained on a first-ever dataset called Brain4Cars including multi-modal inside and outside data of the vehicle.

Ref. [38] introduces an end-to-end learning approach of driving models. Based on visual features of the current and the previous vehicle states, the contribution is the proposition of a holistic approach that predicts the distribution driving behaviors along with semantic segmentation. An adaptive FCN-LSTM model has been proposed to anticipate the distribution of the vehicle motion. A crowd-sourced driving behavior dataset is also introduced to better evaluate the efficiency of the proposed learning paradigm for both semantic segmentation and behaviors prediction.

A more comprehensive approach was introduced in [39] to jointly analyze road agent dynamics, interactions, and the scene context along with the prediction of future states of the road agents, not only the ego vehicle as done by the previously cited work. A CNN architecture was used to encode static and dynamic states and semantic map information in a top-down spatial grid. Due to the lack of a benchmarking dataset that includes road map information, researchers propose a novel large-scale dataset to advance the state-of-the-art of the field.

4. Deep Learning Solutions

4.1. Deep Convolutional Neural Networks

CNNs are the most appealing variant of deep neural networks in vision related-tasks. They were basically developed by LeCun et al. in [40] and successfully reused in the ILSVRC (ImageNet Large Scale Visual Recognition Challenge) competition by Krizhevsky et al. [10]. CNNs are able to perform feature learning to get consistent representation from visual data (basically). Its architecture includes three kinds of layers/mapping functions; convolution, pooling, and ended with one or more fully connected or GAP (Global Average Pooling) layers. Dropout and batch normalization is generally used in CNN for regularization. The biggest advantage is the use of weight sharing (contrary to conventional neural networks), which helps in saving memory and reducing complexity. Along with this, CNNs have shown impressive performance and outperformed all the machine learning techniques devoted to visual understanding.

One of the main application of CNNs is object detection, which represents an important task and a major challenge in computer vision [41]. Its challenges are mainly related to object localization and classification. Several methods based on CNN exist in the literature. Overall, methods using CNNs fall into two categories: (i) single-stage methods, which perform object localization and classification in a single network such as Single-Shot Detector (SSD) [42] and You Only Look Once (YOLO) [43], Both of these architectures produce as output a bounding box of each detected object, its class, and its confidence score [41], and (ii) two-stage methods, which have two separate networks for each of these tasks.

Among its milestones models, the important engines in CNN are AlexNet [10], VGG-16 [44], GoogleNet/Inception [45], ResNet [46], and Xception [47]. Autonomous driving systems take advantage of the so-called pre-trained model of an already-trained model over other visual tasks. Expressly, for a completely new problem/data, the ITS community makes use of the publicly available model to either (i) use open access meta-architectures to train it from scratch on their own data or (ii) use the pre-trained models as feature extractors by feeding new data to a trained model with its trained weights and tweak it to train the new task, which is called transfer learning. Commonly, the output layer is replaced with a new fine-tuned layer in order to determine the deviation of the prediction to the labeled data. Thus, using a non-large-scale dataset, three training ways might be considered: train only the output layer, train the whole CNN, or the last few layers. The ITS community makes use of all these learning strategies to build, train, or retrain their CNN backbones. With application to the on-road behavior tasks, while many endeavors succeed to generate robust features from pre-trained models [6][21][24], major improvements have been proposed recently at many levels of the standard architecture with regards to the spatial exploitation, depth, and feature map exploitation.

Graph Convolutional Network (GCN). GCN consists in using graph-based learning instead of learning data represented in the Euclidean space (such as image or video). The strength of the graph-based framework is the ease of representing the interaction occurring between the instance's components. This is of great importance in road scenes, which will serve in describing the causality between the intention and the attention of the drivers and all the road participants. This paradigm has been exploited by Li et al. in [12] to model interactions, as detailed earlier in Section 4. The two proposed networks, Ego-Thing Graph and Ego-Stuff Graph advanced the state-of-the-art of GCNs through the extension of the backbone proposed in Ref. [48] with two-stream instead of one-stream. First, to generate graphs, 3D convolutions have been applied to obtain the first level of visual features which are feeding both streams; Thing and Stuff graphs. Then, as a second step, the Thing and Stuff representations are extracted using RoIAlign [49] and the newly proposed approach called MaskAlign to deal with irregular objects. The extracted features from both streams are then used to generate the graph generators using a frame-wise fashion. Ego-Thing Graph and Ego-Stuff Graph allow afterward to pass on the connections between different kinds of objects through these GCNs. Outputs from the two obtained streams are fused and fed into a temporal module. This latter module aggregates spatial features using max-pooling to compute the final GCN output.

Fully Convolutional Networks (FCN). Unlike standard CNN, FCN is an end-to-end network where a Fully Connected (FC) layer [50] or an MLP (Multi-Layer Perceptron) network [51] are derived on the top of a CNN-like network where filters are learned at all the levels including the decision-making layers. This allows FCNs to learn representations and scoring based on local features/data. Ref. [38] exploits this fact to propose an end-to-end architecture for generic motion models for autonomous vehicles under crowd contexts. Researers here propose the so-called dilated FCN approach. Taking advantage of the pre-trained CNN models, the second and the fifth pooling layers have been removed and a dilated convolution layer replaced the third convolution layer through FC7. The difference between the usual convolution with stride and the dilated one is the expansion of the filter's size before doing the convolution.

Among other applications of FCN is semantic segmentation, where the output of the model has the same resolution as the input images, with a class prediction for each pixel. Since the pioneering work of Long et al. [52], where the base structure of the model is an encoder and a decoder streams, many variants have been proposed. Most of these works focus on the improvement of the accuracy of segmentation. However, real-time performance is very important for autonomous driving. Combination of light architectures like SkipNet [52] and ShuffleNet [53] for the encoder and decoder parts, respectively, allows segmentation rates at about 16 FPS on a Jetson TX2, while maintaining high accuracy [54].

3D CNNs. represent an extension of CNNs where a 3D activation map is generated over the convolution layers. The intuition behind this is to encode data that can be represented on more than two dimensions like volumetric or temporal data. 3D CNNs have been used in different contexts such as 3D shape estimation [55], human activity detection [56], and recently explored in RBA systems [57] to further enhance the understanding of drivers' behaviors. Indeed, a TRB (Temporal Reasoning Block) has been introduced to model the causes of behaviors. The aim of this 3D-CNN is to discriminate spatio-temporal representations with attention saliency mechanisms. By assuming inputs as coarse-grained

videos, the main contribution here is the proposition of a novel reasoning block composed of two layers; the first one consists of fine-grained 3D convolution and the second one allows to keep the temporal continuity.

CNN Features Extractors. With the availability of large-scale visual data along with optimization algorithms and powerful CPUs/GPUs, it becomes possible to train deep networks that achieve impressive performance on roughly all the challenging tasks. The obtained models are shared for the community in order to avoid training from scratch and make it easier for researchers to enhance the available models or reuse them as features extractors. For example, a CNN network that has been trained to classify road objects will output several features from the low-level to the high-level layers with increasing complexity and abstraction. Complexity, in this case, goes from pixels, blobs, circles, wheels, stuff, faces, hoods until bicycles, cars, pedestrians, and the whole scene. A number of neurons are supposed to be activated for these abstraction levels. The key feature of this is that another classification task devoted for example for the indoor or in-the-wild scenes will make use of the same low- and mid-level features that are present in all the domains.

The ImageNet [10] and the COCO [58] pre-trained models are widely considered for RBA tasks and reused as backbones for feature extraction. The (meta-)architectures of neural networks did make use of state-of-the-art backbones along with task-specific layers (for, e.g., classification or detection heads). Thus, choosing the right feature extractor is important since its properties (mainly the type of layers and number of parameters) directly affect the performance of the whole network. By barring the few examples that decouple the backbone from the meta-architecture [59], the main feature extractors used in the related works are AlexNet [10], ResNet-50 [46], VGG-16 [44], Inception-v2 [60], and InceptionResnet-v2 [61] used in the following non-exhaustive list of papers detailed above [5][6][9][21][31], respectively.

4.2. Deep Recurrent Neural Networks

Despite its impressive performance for tasks related to image analysis, CNNs examine only the current input and fail in handling sequential data. RNNs (Recurrent Neural Networks) however process only sequential data [62]. Composed often of a single node with internal memory, the principle of RNNs is memorizing the outputs and feeding them back as inputs and continuing to do this until predicting the output of the layer. Thus, RNNs allow saving information that has occurred in the past and looks for patterns over time and the length of the sequence. With application to ADAS and AV systems, such end-to-end machines have been recently used to automatically model non-linear discriminative representations to improve the performance of vision-based analysis tools basically related to RBA.

To start, an LSTM network has been used in Ref. [6] to propose the so-called ADMD system composed of four blocks to learn driving maneuvers as spatio-temporal sequences. Features that serve as an input for the predictor, i.e., LSTM network, were transferred from a CNN model (InceptionResnet-v2 [61]), as in [11], an attention map generator, and raw vehicle signals.

Similarly, a novel network architecture called TraPHic is introduced in [28] to predict road agent trajectories under complex contexts. The issue with LSTM under such circumstances is the inability to model relationships of heterogeneous road agents since the settings of an LSTM unit are independent of the others. To capture temporal dependencies of objects' spatial coordinates, LSTMs are employed and combined with CNN to boost the learning of local objects' relationships in space and time. In Ref. [31], a two-stream deep architecture for event proposal and prediction is proposed. The first one proposed the key-frames of the event sequences. A standard LSTM classifier is employed to predict the corresponding class among approaching, entering, passing. The output of this first stream, the candidate frames, is then forwarded to the second stream of prediction. The frames are here aggregated through GAP and output the event class.

As for the GRU networks applied for RBA tasks, this is still a niche area. For instance, Hong et al. [39] propose a new encoder-decoder architecture where an RNN-based composed of a single GRU cell has been employed as a decoder. Unlike LSTM units, GRU has the ability to control the inputs without memorizing in-between status. This allows a less complex decoder block in this encoder-decoder architecture [39]. A more fancy unit has been proposed in Ref. [32] called GRFU (Gated Recurrent Fusion Units) aiming to learn temporal data and fusion simultaneously. Precisely, the novel gating mechanism learns a representation of every single-mode, i.e., sensor, for each instance in order to infer the best fusion strategy among Late Recurrent Summation (LRS), Early Gated Recurrent Fusion (EGRF), or Late Gated Recurrent State Fusion (LGRF).

References

1. Xique, I.J.; Buller, W.; Fard, Z.B.; Dennis, E.; Hart, B. Evaluating Complementary Strengths and Weaknesses of ADAS Sensors. In Proceedings of the 2018 IEEE 88th Vehicular Technology Conference (VTC-Fall), Chicago, IL, USA, 27–30

August 2018; pp. 1–5.

2. Hu, H.N.; Cai, Q.Z.; Wang, D.; Lin, J.; Sun, M.; Krahenbuhl, P.; Darrell, T.; Yu, F. Joint Monocular 3D Vehicle Detection and Tracking. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 5390–5399.
3. Bresson, G.; Alsayed, Z.; Yu, L.; Glaser, S. Simultaneous localization and mapping: A survey of current trends in autonomous driving. IEEE Trans. Intell. Veh. 2017, 2, 194–220.
4. Santhosh, K.K.; Dogra, D.P.; Roy, P.P. Temporal unknown incremental clustering model for analysis of traffic surveillance videos. IEEE Trans. Intell. Transp. Syst. 2018, 20, 1762–1773.
5. Yan, X.; Wang, F.; Liu, W.; Yu, Y.; He, S.; Pan, J. Visualizing the Invisible: Occluded Vehicle Segmentation and Recovery. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 7618–7627.
6. Peng, X.; Zhao, A.; Wang, S.; Murphey, Y.L.; Li, Y. Attention-Driven Driving Maneuver Detection System. In Proceedings of the 2019 International Joint Conference on Neural Networks (IJCNN), Budapest, Hungary, 14–19 July 2019; pp. 1–8.
7. Kooij, J.F.P.; Schneider, N.; Flohr, F.; Gavrila, D.M. Context-based pedestrian path prediction. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; Springer: Berlin/Heidelberg, Germany, 2014; pp. 618–633.
8. Meinecke, M.M.; Obojski, M.; Gavrila, D.; Marc, E.; Morris, R.; Tons, M.; Letellier, L. Strategies in terms of vulnerable road user protection. EU Proj. SAVE-U Deliv. D 2003, 6, 2003.
9. Rasouli, A.; Kotseruba, I.; Tsotsos, J.K. Are they going to cross? A benchmark dataset and baseline for pedestrian crosswalk behavior. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 206–213.
10. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. In Proceedings of the Advances in Neural Information Processing Systems, Lake Tahoe, NV, USA, 3–8 December 2012; pp. 1097–1105.
11. Ramanishka, V.; Chen, Y.T.; Misu, T.; Saenko, K. Toward driving scene understanding: A dataset for learning driver behavior and causal reasoning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Istanbul, Turkey, 30–31 January 2018; pp. 7699–7707.
12. Li, C.; Meng, Y.; Chan, S.H.; Chen, Y.T. Learning 3D-aware Egocentric Spatial-Temporal Interaction via Graph Convolutional Networks. arXiv 2019, arXiv:1909.09272.
13. Wu, J.; Wang, L.; Wang, L.; Guo, J.; Wu, G. Learning Actor Relation Graphs for Group Activity Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Seoul, Korea, 27 October–2 November 2019; pp. 9964–9974.
14. Oliver, N.; Pentland, A.P. Graphical models for driver behavior recognition in a smartcar. In Proceedings of the IEEE Intelligent Vehicles Symposium 2000 (Cat. No. 00TH8511), Dearborn, MI, USA, 5 October 2000; pp. 7–12.
15. Singh, D.; Mohan, C.K. Deep spatio-temporal representation for detection of road accidents using stacked autoencoder. IEEE Trans. Intell. Transp. Syst. 2018, 20, 879–887.
16. Hu, X.; Xu, X.; Xiao, Y.; Chen, H.; He, S.; Qin, J.; Heng, P.A. SINet: A scale-insensitive convolutional neural network for fast vehicle detection. IEEE Trans. Intell. Transp. Syst. 2018, 20, 1010–1019.
17. Chen, B.; Gong, C.; Yang, J. Importance-aware semantic segmentation for autonomous vehicles. IEEE Trans. Intell. Transp. Syst. 2018, 20, 137–148.
18. Aly, M. Real time detection of lane markers in urban streets. In Proceedings of the 2008 IEEE Intelligent Vehicles Symposium, Eindhoven, The Netherlands, 4–6 June 2008; pp. 7–12.
19. Gong, J.; Jiang, Y.; Xiong, G.; Guan, C.; Tao, G.; Chen, H. The recognition and tracking of traffic lights based on color segmentation and camshift for intelligent vehicles. In Proceedings of the 2010 IEEE Intelligent Vehicles Symposium, San Diego, CA, USA, 21–24 June 2010; pp. 431–435.
20. Gavrila, D.M.; Giebel, J.; Munder, S. Vision-based pedestrian detection: The PROTECTOR system. In Proceedings of the IEEE Intelligent Vehicles Symposium, Parma, Italy, 14–17 June 2004; pp. 13–18.
21. Zhou, W.; Lv, S.; Jiang, Q.; Yu, L. Deep Road Scene Understanding. IEEE Signal Process. Lett. 2019, 26, 587–591.
22. Fauqueur, J.; Brostow, G.; Cipolla, R. Assisted video object labeling by joint tracking of regions and keypoints. In Proceedings of the 2007 IEEE 11th International Conference on Computer Vision, Rio de Janeiro, Brazil, 14–20 October 2007; pp. 1–7.

23. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Trans. Pattern Anal. Mach. Intell. 2017, 40, 834–848.

24. Brust, C.A.; Sickert, S.; Simon, M.; Rodner, E.; Denzler, J. Convolutional Patch Networks with Spatial Prior for Road Detection and Urban Scene Understanding. arXiv 2015, arXiv:1502.06344.

25. Geiger, A.; Lenz, P.; Urtasun, R. Are we ready for autonomous driving? The kitti vision benchmark suite. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 18–20 June 2012; pp. 3354–3361.

26. Frohlich, B.; Rodner, E.; Denzler, J. A fast approach for pixelwise labeling of facade images. In Proceedings of the 2010 20th International Conference on Pattern Recognition, Istanbul, Turkey, 23–26 August 2010; pp. 3029–3032.

27. Choi, C.; Patil, A.; Malla, S. Drogon: A causal reasoning framework for future trajectory forecast. arXiv 2019, arXiv:1908.00024.

28. Chandra, R.; Bhattacharya, U.; Bera, A.; Manocha, D. Traphic: Trajectory prediction in dense and heterogeneous traffic using weighted interactions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 8483–8492.

29. Chandra, R.; Bhattacharya, U.; Mittal, T.; Li, X.; Bera, A.; Manocha, D. GraphRQI: Classifying Driver Behaviors Using Graph Spectrums. arXiv 2019, arXiv:1910.00049.

30. Chang, M.F.; Lambert, J.; Sangkloy, P.; Singh, J.; Bak, S.; Hartnett, A.; Wang, D.; Carr, P.; Lucey, S.; Ramanan, D.; et al. Argoverse: 3D Tracking and Forecasting with Rich Maps. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 8748–8757.

31. Narayanan, A.; Dwivedi, I.; Dariush, B. Dynamic Traffic Scene Classification with Space-Time Coherence. arXiv 2019, arXiv:1905.12708.

32. Narayanan, A.; Siravuru, A.; Dariush, B. Temporal Multimodal Fusion for Driver Behavior Prediction Tasks using Gated Recurrent Fusion Units. arXiv 2019, arXiv:1910.00628.

33. Taha, A.; Chen, Y.T.; Misu, T.; Davis, L. In Defense of the Triplet Loss for Visual Recognition. arXiv 2019, arXiv:1901.08616.

34. Guangyu Li, M.; Jiang, B.; Che, Z.; Shi, X.; Liu, M.; Meng, Y.; Ye, J.; Liu, Y. DBUS: Human Driving Behavior Understanding System. In Proceedings of the IEEE International Conference on Computer Vision Workshops, Seoul, Korea, 27 October–2 November 2019.

35. Chen, Y.; Wang, J.; Li, J.; Lu, C.; Luo, Z.; Han, X.; Wang, C. LiDAR-Video Driving Dataset: Learning Driving Policies Effectively. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 5870–5878.

36. Jain, A.; Koppula, H.S.; Raghavan, B.; Soh, S.; Saxena, A. Car that knows before you do: Anticipating maneuvers via learning temporal driving models. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 3182–3190.

37. Bengio, Y.; Frasconi, P. An input output HMM architecture. In Proceedings of the Advances in Neural Information Processing Systems, Denver, CO, USA, 27–30 November 1995; pp. 427–434.

38. Xu, H.; Gao, Y.; Yu, F.; Darrell, T. End-to-end learning of driving models from large-scale video datasets. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–27 July 2017; pp. 2174–2182.

39. Hong, J.; Sapp, B.; Philbin, J. Rules of the Road: Predicting Driving Behavior with a Convolutional Model of Semantic Interactions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 8454–8462.

40. LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. Proc. IEEE 1998, 86, 2278–2324.

41. Mauri, A.; Khemmar, R.; Decoux, B.; Ragot, N.; Rossi, R.; Trabelsi, R.; Boutteau, R.; Ertaud, J.Y.; Savatier, X. Deep learning for real-time 3D multi-object detection, localisation, and tracking: Application to smart mobility. Sensors 2020, 20, 532.

42. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. Ssd: Single shot multibox detector. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; Springer: Berlin/Heidelberg, Germany, 2016; pp. 21–37.

43. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 779–788.
44. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. arXiv 2014, arXiv:1409.1556.
45. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–15 June 2015; pp. 1–9.
46. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
47. Chollet, F. Xception: Deep learning with depthwise separable convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1251–1258.
48. Kipf, T.N.; Welling, M. Semi-Supervised Classification with Graph Convolutional Networks. In Proceedings of the International Conference on Learning Representations (ICLR), Toulon, France, 24–26 April 2017.
49. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2961–2969.
50. Dai, J.; Li, Y.; He, K.; Sun, J. R-fcn: Object detection via region-based fully convolutional networks. In Proceedings of the Advances in Neural Information Processing Systems, Barcelona, Spain, 5–10 December 2016; pp. 379–387.
51. Maggiori, E.; Tarabalka, Y.; Charpiat, G.; Alliez, P. Can semantic labeling methods generalize to any city? The inria aerial image labeling benchmark. In Proceedings of the 2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Worth, TX, USA, 23–28 July 2017; pp. 3226–3229.
52. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
53. Zhang, X.; Zhou, X.; Lin, M.; Sun, J. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 6848–6856.
54. Siam, M.; Gamal, M.; Abdel-Razek, M.; Yogamani, S.; Jagersand, M.; Zhang, H. A comparative study of real-time semantic segmentation for autonomous driving. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Salt Lake City, UT, USA, 18–22 June 2018; pp. 587–597.
55. Yi, L.; Su, H.; Guo, X.; Guibas, L.J. Syncspeccnn: Synchronized spectral cnn for 3d shape segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2282–2290.
56. Xu, H.; Das, A.; Saenko, K. R-c3d: Region convolutional 3d network for temporal activity detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 5783–5792.
57. Liu, Y.C.; Hsieh, Y.A.; Chen, M.H.; Yang, C.H.H.; Tegner, J.; Tsai, Y.C.J. Interpretable Self-Attention Temporal Reasoning for Driving Behavior Understanding. arXiv 2019, arXiv:1911.02172.
58. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; Springer: Berlin/Heidelberg, Germany, 2014; pp. 740–755.
59. Lee, S.; Kim, J.; Oh, T.H.; Jeong, Y.; Yoo, D.; Lin, S.; Kweon, I.S. Visuomotor Understanding for Representation Learning of Driving Scenes. arXiv 2019, arXiv:1909.06979.
60. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the inception architecture for computer vision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 2818–2826.
61. Szegedy, C.; Ioffe, S.; Vanhoucke, V.; Alemi, A.A. Inception-v4, inception-resnet and the impact of residual connections on learning. In Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017; pp. 4278–4284.
62. Lipton, Z.C.; Berkowitz, J.; Elkan, C. A critical review of recurrent neural networks for sequence learning. arXiv 2015, arXiv:1506.00019.