26th ITS World Congress, Singapore, 21-25 October 2019

Paper ID # EU-TP2296

Scene danger ranking using deep neural network

Remy Bendahan[1*], Dzmitry Tsishkou[2], Frederic Abad[1]

1. IMRA EUROPE SAS, 220 Rue Albert Caquot, B.P. 213, 06904 Sophia-Antipolis, France, e-mail: name@imra-europe.com

2. was working at IMRA Europe S.A.S. during this R&D activity

Abstract

Next generation of automotive driving support requires the ability to anticipate driving hazards better than humans. We present a novel approach for "copying" and aggregating driving expert knowledge of scene danger ranking by using a multi-task DNN. As every danger is different and that it is impossible to gather all examples of danger in a training dataset, we inspired from humans and showed that combining features from obstacles, motion, distance, possible trajectories, focus and anticipation allows to reason about the context of new scenes to rank the danger. We established an incremental transfer learning process for training from doubly sparse labels (few samples of hazard represented by few pixels). We present the danger dataset and provide an analysis of the ranking ability. From one pixel, the DNN is able to provide the danger level for every pixel in the image. We conclude positively about the feasibility of scene danger ranking.

Keywords: **Danger Level, Transfer Learning, Learning from sparse labels**

Introduction

Next generation of automotive driving support (e.g. connected autonomous vehicles - CAV) will certainly require the ability to anticipate driving hazards better than humans. For instance, to avoid hazardous situations and navigate safely in crowded urban scenario, CAV must realize complex judgements such as the ranking of the dangerous areas in a scene. Our aim is to create a machine with the skill of anticipating what will happen on the road ahead by considering complex hazardous situations.

We consider that the key for safe maneuver decisions of CAV is to use the absolute danger level of each area of the scene as input to path planning algorithms. Whatever the driving scenario is, this information will in particular allow to dynamically adjust the margin around obstacles according to their level of danger. As opposition of today's fixed navigation margins [3], dynamic margins would avoid deadlock of CAV and allow continuity of service in narrow and crowded space.

Today's path planning decisions are computed from dynamic maps by combining high accuracy 3D NAVI map, ego-vehicle dynamics data and obstacle map made from multiple sensors to detect obstacles and predict their positions based on their dynamics [5], [6]. Path planning algorithms could be enhanced by adding the danger map to the dynamic map (Figure 1), which would insure minimal

risk during navigation (safest path).

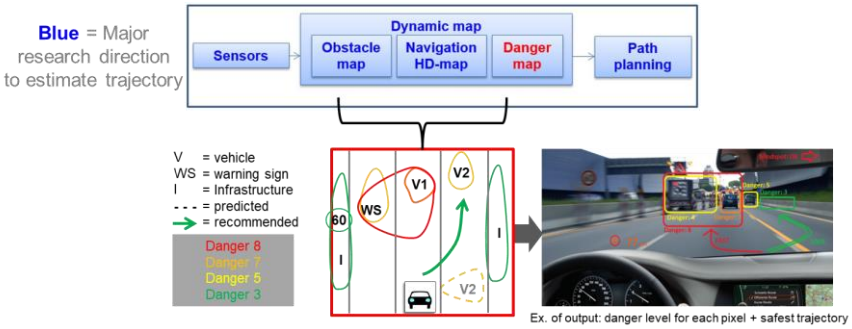

Figure 1 – Danger Level Map allows the computation of the safest path that anticipates possible risks and decides their level

In this paper, we will present our approach for "copying" and aggregating driving expert knowledge of scene danger ranking by using deep neural network (DNN). In section 'Related works', we will review prior art on danger estimation and position our contributions. In section 'End-to-End DNN for scene danger ranking', we will describe and justify how we aim to cover all dangers (felt by humans) in every scene in a comprehensive way. We will show how we plan to cope with the facts that every danger is different and that it is impossible to gather all examples of danger in a training dataset. As real hazards are much less frequent that non-hazardous situations, we will show how we cope for training a DNN from doubly sparse labels (few samples of hazard represented by few pixels in input images). This is done by incremental transfer learning of multi tasks inspired from human reasoning. In section 'Experiments', we will present how we gathered the most important dataset: the dataset of danger levels. Next, we will describe in details the DNN architecture and the training process. We will discuss the achievements and explain why we concluded positively about the feasibility of absolute scene danger ranking from knowledge captured based on subjective danger level appraisal. Finally, we will conclude and propose the next actions to demonstrate large-scale danger raking ability of a DNN.

Related works on danger and risk assessment

In [11] the authors predict generic (not specifically driving) accidents in an agent-centric way using a soft-attention Recurrent Neural Network (RNN) modelling the interactions between the agents or static regions involved in the accidents. This system cannot cope with sparse labels, cannot estimate levels of danger when there is no danger event explicitly happening and it cannot handle situations with simultaneous multiple potentially dangerous objects.

In [4], the authors calculate the level of danger of vehicle collision with in-path pedestrians based on pedestrian distance, response distance and braking distance. In [10] the authors observe that irregular motion behavior and low illumination are additionally sources of threat. These systems strongly assume that obstacle detection is always successful and either rely on limited types of dangerous scene elements or limited danger clues.

In [1], the authors propose a method to identify the current Traffic Scene (TS) from a multiplicity of modalities and to classify this TS as dangerous by using a neural network trained on TS labelled as

dangerous using predefined rules. This system is limited to recognize situations inside the range of finite rules hard-coded by humans. Additionally, this system cannot identify the scene elements that make the situation dangerous which is necessary for safest path computation.
In [9], the authors design a Markov Chain model to predict the driving risk status from instantaneous driving risk levels determined in time-to-collision (ttc) and time-headway (thd) two-dimension plane and feature vectors made from information such as vehicle movement, traffic, environmental status and obstacle detection. This model is based on high level information, and does not use any intermediate information coming from the input of imaging sensors (LIDAR, camera). Its prediction is therefore blind to events that are not taken into account by the vector features (crossing animal, flare blinding the driver…) and do not localize hazards.

From the literature, we can understand that an ideal danger prediction model should be able to handle sparse labels and situations where the danger is not yet developing or when multiple elements in the scene are potentially dangerous. The model should not be limited by hard-coded rules and should identify all the hazardous elements in the scene. Finally, the model should not depend explicitly on obstacle detection and should instead benefit from all features available from the imaging sensors.

End-to-End DNN for scene danger ranking

As previously stated, each hazardous situation is unique but yet humans (drivers) have developed the ability to reason about the context of the scene to safely drive in each new situation. In [7], the authors analyze how drivers detect and respond to roadway hazards and propose a framework of hazard avoidance. They introduce the definitions of hazards, their precursors, the prioritization of precursors. Monitoring of hazards is described as resulting from overt and covert attention mechanisms. Indeed, while facing a hazardous situation (during driving), human drivers cannot explain how they decide about hazard locations, hazard priorities and vehicle control. It confirms that prioritizing hazard is a covert mechanism and thus there is no perfect objective measure of hazard (localization and priority level). The danger ranking DNN must reproduce subjectively appraised danger locations and levels.

Drivers (experts) are however better to explain their subjective decisions a posteriori (e.g. by watching recorded videos). Questionnaires could have been useful to gather the required training dataset. But unfortunately, we could never gather enough hazard samples to insure high reliability in training end-to-end DNN using solely danger locations and levels as image labels. This is why we propose a DNN architecture that combines the information used by human for reasoning about hazards.

DNN architecture inspired from human reasoning

When expert drivers are asked to explain a posteriori why they decided to react to hazards, several keywords appear relevant: **obstacles, motion, distance, possible trajectories, focus and hazard anticipation**. These are key concepts that most probably allow humans to reason over scenes in order to successfully handle new situations. But the state of art in driving psychology and behavior analysis

[7] does not explain how to efficiently combine this information to decide about hazards.

We aim to provide end-to-end multi task deep neural network that takes an image or a time series of images as input, generates the above-mentioned relevant information, and uses them to generate a pixel-wise danger ranking image as output (Figure 1). Major arguments towards end-to-end learning are: first, by learning a deep neural network end-to-end we could optimize every neural connection at once, which is the most efficient way of training (especially when we do not know how to combine key information); second, end-to-end learning is adapted to modelling abstract concepts (e.g. danger, its ranking and its anticipation); and third, the end-to-end trained neural network is easily transferable and compatible with all deep learning frameworks, so our danger ranking deep neural network could be easily installed as a building block of any automotive systems.

Training via incremental transfer learning

The danger ranking training method we are proposing enables to train from sparse labels a deep neural network to identify within images or image sequences the regions of various danger levels. We have designed a specific deep transfer learning [8] curriculum which both compensates for the sparsity of training data and ensures that our system will learn and take advantage of the multiple key concepts used by expert drivers' reasoning. Such achievement is possible thanks to the decomposition of the training procedure into several training steps including generic (A) and specific image recognition (B), specific motion and distance estimation (C), object trajectory prediction (D) and focus on potential danger areas (E) acquired through saliency and eye tracking. The process we have designed, illustrated in Figure 2, starts by training blocks to learn generic knowledge with larger amount of training data and next transfers the learning to following blocks learning more specific knowledge with not enough training samples to be learned as standalone processes.

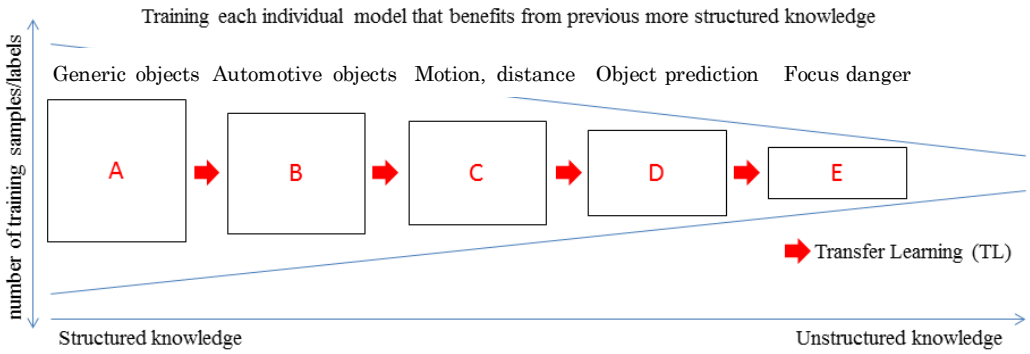

Figure 2: Flow chart of transfer knowledge learning

The two object recognition blocks (A) and (B) allow the training method to first learn the features necessary for classifying any object and next classifying specific objects related to the automotive context. Distance and motion estimations (C) ensure learning the features to understand how the objects are positioned with respect to the ego-vehicle and how they move in the environment. Object trajectory prediction (D) allows learning how objects will move with respect to each other (chain

effect). Saliency-based danger focusing (E) trains the method for learning how humans react to new hazardous environment. Finally, the end-to-end deep neural network (combined knowledge - F) includes the top layers of the transfer learning blocks, thus receiving the trained features associated to each previous knowledge as shown in Figure 3. As can be seen in Figure 2 and Figure 3, the transfer learning is incrementally applied to transfer the most structured (object-wise) and dense knowledge (e.g. A and B) to the less structured (saliency, danger) and sparse knowledge (e.g. E, F).

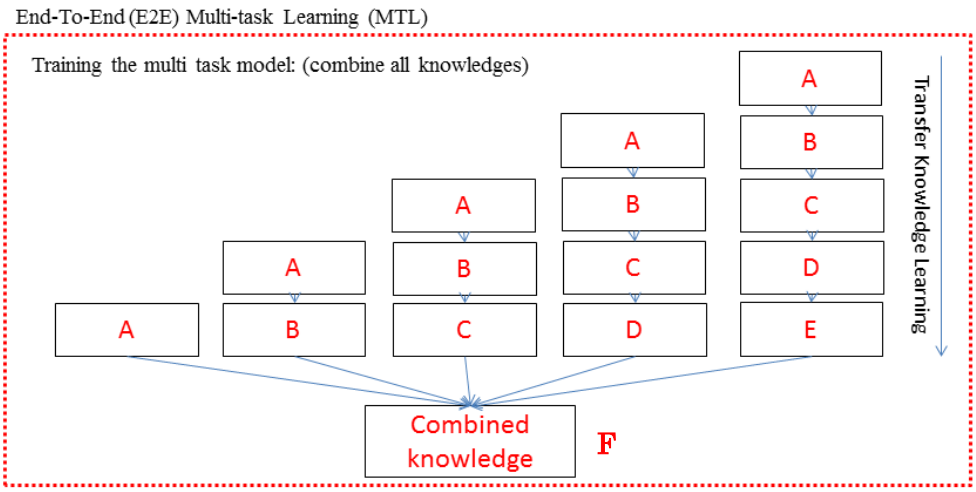

Figure 3: Flow chart for knowledge combination.

Experiments

In our experiments, we have used the transfer learning process as described above. The combined knowledge (DNN-F in Figure 3) corresponds to the final end-to-end DNN that could optimally combine the different information to finally evaluate for each pixel, the danger location and its level. This DNN plays a role of integrator or "chef d'orchestre".

In order to perform series of transfer learning steps one must use similar architectures of DNN, so that each consecutive training starts from a previously trained model and fine-tune it by using a new dataset specific to the new knowledge we want to acquire. Whereas the datasets required to train A~E blocks are rather standard datasets, the dataset required to train the DNN for the 'combined knowledge F' block needs to be labelled with danger location and danger level. This dataset and the final architecture are presented in the next sections.

Dataset of driving situation labelled with sparse labels of danger levels

The dataset was constructed with the goal to test the feasibility of creating a model for danger localization and ranking using very sparse labels. We agreed on specific criteria of danger definition. The subject who annotated the dataset consistently was looking at still images (looking at video footage will be done in next experiment). The subject was requested to actively search the area with maximum danger level according to the following definition:

- Danger Level-0: no action needed to change ego-car trajectory

- Danger Level-1: anticipatory braking or lane change when collision risk is minimal
- Danger Level-2: controlled braking or lane change with enough time to perform maneuver
- Danger Level-3: rapid braking or lane change or stopping required to avoid collision
- Danger Level-4: emergency braking or violent steering to avoid collision resulting in near miss or eventually a mitigated collision

Current dataset is composed of original cityscapes (3.5K) images (train+val) [2]. The detailed protocol for danger rating was the followings:

- The subject (16 years experimented driver with no accident record) sees a black screen for 250ms (time to forget previous image taken from typical psychology experiments)
- The subject observes the image (640 x 320) up to 1500ms
- The subject clicks with mouse pointer on the object that he considers the most dangerous
- System records the location of the pixels and subjects's reaction time (x, y, reaction time)
- System converts single pixel to a circle of diameter 5% of image (average object size) width
- Danger label is created with Rank of danger between 0-4 (low-high) related to reaction time (T): rank_danger= 5*(T-450)/1100, where 450ms is the minimum reaction time measured experimentally and 1100ms is normalized corresponding the maximum reaction time of 1500ms measured experimentally.

In this experiment, the danger level is proportional to the time required by the subject to see the dangerous area. The longer the subject searches in the image, the more difficult it is to decide the danger level and the more dangerous/hazardous the situation is. Totally 5 sessions of labelling were made during 2 months of tagging by the same subject. Due to the fact that the same person was used, we expect coherency between labels, danger levels and scenes. This coherency is particularly useful to get absolute danger ranking which will allow the DNN to treat all scenes in the same way. We consider these labels as a good way to capture coherently the subjectivity of self-rating of danger levels. These labels should be a good start for our feasibility study but in the future, they should be complemented and validated by multiple experts such as driving instructors. This will allow to further rely on the input training data and also aggregate the multiple knowledges to finally get a DNN better than a single expert.

Detailed training procedure

A new architecture of deep neural network FCN8s-Pyramid was proposed. It uses VGG16 for convolutional part and our own design of de-convolutional part inspired by state of art PSP-Network [12]. The latter part was modified to replace the original ResNet base, for compatibility and training acceleration purposes.

As shown in Figure 4, our multi-task deep neural network architecture (green) shares a single VGG16 convolutional part (blue – block A) trained on ImageNet 1M dataset and combines into one fusion block (green)5 deconvolution parts (input image size 384x384x3, limited by 12GB size of RAM on

modern GPU) each corresponding to a specific transfer learning block (red - B~E). At the time being the block (D) of object prediction is not yet integrated in the architecture. To cope with this lack, we gave during training to the DNN F the access to present and future frames which implicitly include the information about future trajectories.

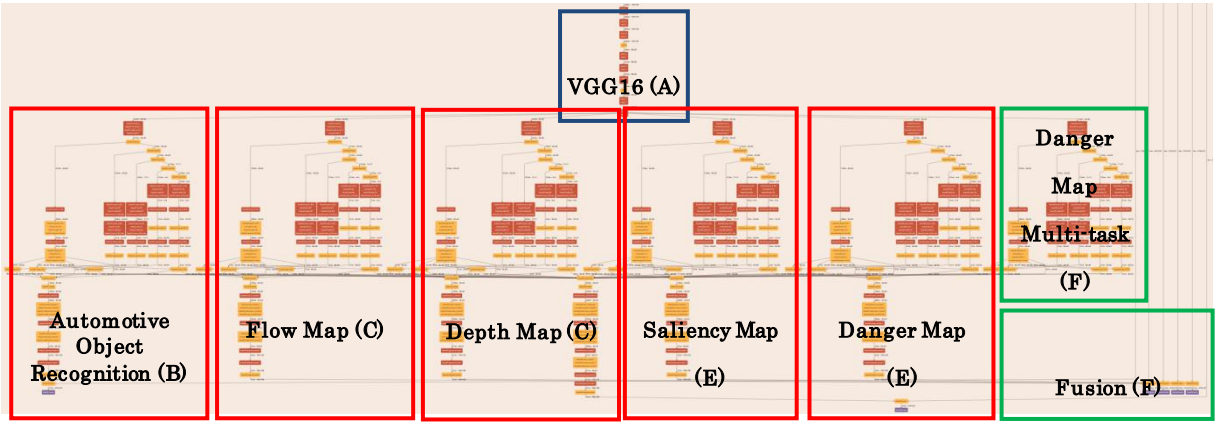

Figure 4: Network architecture for multi-task fusion of 5 transfer learning blocks.

The final multi-task danger ranking module (F) that aggregates all transfer learning blocks was trained using very sparse labels as described in the previous section. Using middle level GPU (GTX 1060) of 6GB of RAM, each block was individually trained with a task specific dataset (public domain datasets) and fine-tuned on own recorded videos. We could complete the training in 1 week and achieved 10 fps for inference processing.

Experimental results and discussion

We aimed to confirm the merit of combining the different information for the task of danger ranking. The visual interface to analyze live danger level maps and those of transfer learning blocks is shown on Figure 5. Activation maps show the importance of image area (red=high level, blue=low level, black=non-important) for each transfer learning block for final decision of danger rank for each pixel.

Figure 5 Results of danger ranking and transfer learning blocks (low danger levels not shown)

The analysis of activation maps demonstrates how differently each transfer learning block focuses on the most important image area from its own perspective:

- Automotive object recognition (B): mostly focused on objects on the right, including main focus on nearby pedestrians. We hypothesize that this block is directly looking for objects.
- Depth map (C): major focus on the road center. We hypothesize that this block estimates free space.
- Saliency map (E): major focus on horizon. We hypothesize that this block aims to detect newly appearing object in long range (in case of empty road).
- Danger ranking single-task (E): focused mostly on close road area. We hypothesize that this block is looking for nearby obstacles on a road.
- Danger ranking multi-task (F): very narrow but high-level focus on results of above blocks. We hypothesize that this block plays a role of integrator of information and ranking by priority (benefit of multi-task compared to single-task)

Our architecture was confirmed as suitable with middle level GPUs and able of solving the multi-task fusion at 10fps. Qualitative analysis of multi-task danger ranking DNN performance in various driving scenarios suggests that danger ranking benefits from transfer learning. This supports our main hypothesis that danger ranking is feasible via sequence of transfer learning from multiple tasks.

The qualitative analysis of the danger ranking (F) on scene that were never used during the training, showed that it is possible to realize the task of danger ranking. Danger ranking DNN (F) was able to rank imminent danger and interestingly to also rank hidden danger (such as between 2 parked cars Figure 6 - b). At present, we consider that the danger ranking task is not completed. For example it is not clear to us why the pedestrian moving along the pavement is ranked as dangerous (Figure 6 - e)

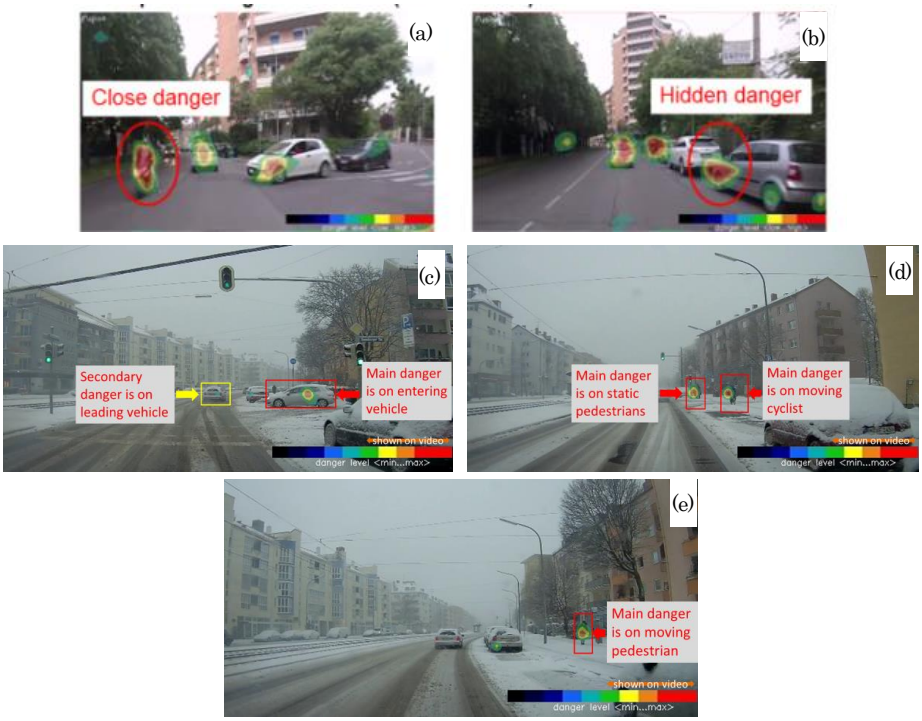

Figure 6 Example of results of danger ranking – (a,b,c,d) expected, (e) unexplained behavior.

Note that at present, we cannot provide quantitative figures on the performance because we have not defined yet how to efficiently evaluate 1 pixel click ground truth with the resulting dense pixel map.

Conclusion

- This paper showed that it is feasible to reproduce by DNN subjectively rated information such as danger ranking. From our knowledge, it is the first DNN of its kind. It opens the possibility to drastically enhance the maneuver planning algorithms to be integrated in CAV.
- Qualitative analysis of multi-task danger ranking performance in various driving scenarios suggests that danger ranking benefits from transfer learning of information inspired from human reasoning.
- Initial version of danger ranking via multi-task fusion of transfer learning blocks was realized. FCN8s-Pyramid architecture was confirmed as suitable to be used with middle level GPUs and capable of solving the multi-task fusion of different blocks.
- The training method allowed us to confirm that dense danger map (each pixel of image) could be obtained from sparse input. This suggests that creating a training dataset of single pixel danger levels annotation per frame would be sufficient to rank the dangers, thus making this method scalable.
- We do not claim that danger ranking task is solved. We have demonstrated its feasibility. Despite the various tasks that are pending (integrate trajectory prediction, validate the labels of danger levels and aggregate more labels from many experts, process video instead of single frames), the major challenge that we are facing is to build a multi-disciplinary team combining multiple knowledges (deep learning, cognitive science, driving psychology, vehicle control) to strengthen our scientific choices.
- We are open to collaborations in this challenging domain, to make it happen…

References

1. Cosatto E., Melving I. and Graf H. P. (2017). Multi-Modal Driving Danger Prediction System for Automobiles, NEC Laboratories America Inc., in Patent US2017023837A1. (link)
2. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., ... & Schiele, B. (2016). The cityscapes dataset for semantic urban scene understanding. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3213-3223).
3. Fairfield N, Furman V. (2017). Testing predictions for autonomous vehicles. Waymo Llc., in Patent WO2018009391A1. (link)
4. García, F., Escalera, A. D. L., Armingol, J. M., García, J., & Llinas, J. (2011). Fusion based safety application for pedestrian detection with danger estimation. In Proceedings of the 14th International Conference on Information Fusion (FUSION 2011). Chicago, Illinois, USA. IEEE. (link)

5. Katrakazas, C., Quddus, M., Chen, W. H., & Deka, L. (2015). Real-time motion planning methods for autonomous on-road driving: State-of-the-art and future research directions. Transportation Research Part C: Emerging Technologies, 60, 416-442.

6. Paden, B., Čáp, M., Yong, S. Z., Yershov, D., & Frazzoli, E. (2016). A survey of motion planning and control techniques for self-driving urban vehicles. IEEE Transactions on intelligent vehicles, 1(1), 33-55.

7. Pradhan A., Crundall D. (2016). "Hazard Avoidance in Young Novice Drivers: Definitions and a Framework", Chapter 6 in "Handbook of Teen and Novice Drivers: Research, Practice, Policy, and Directions". CRC Press.

8. Tan, C., Sun, F., Kong, T., Zhang, W., Yang, C., & Liu, C. (2018, October). A Survey on Deep Transfer Learning. In International Conference on Artificial Neural Networks (pp. 270-279). Springer, Cham.

9. Xiong, X., Chen, L., & Liang, J. (2018). Vehicle Driving Risk Prediction Based on Markov Chain Model. Discrete Dynamics in Nature and Society, 2018. (link)

10. Yuan, Y., Fang, J., & Wang, Q. (2018). Incrementally perceiving hazards in driving. Neurocomputing, 282, 202-217. (link)

11. Zeng, K. H., Chou, S. H., Chan, F. H., Carlos Niebles, J., & Sun, M. (2017). Agent-centric risk assessment: Accident anticipation and risky region localization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 2222-2230). (link)

12. Zhao, H., Shi, J., Qi, X., Wang, X., & Jia, J. (2017, July). Pyramid scene parsing network. In IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) (pp. 2881-2890).