Pyramid Bayesian Method for Model Uncertainty Evaluation of Semantic Segmentation in Autonomous Driving

Yang Zhao¹ · Wei Tian¹ · Hong Cheng¹

Received: 3 November 2020 / Accepted: 14 October 2021 / Published online: 14 January 2022
© The Author(s) 2021

Abstract

With the fast-developing deep learning models in the field of autonomous driving, the research on the uncertainty estimation of deep learning models has also prevailed. Herein, a pyramid Bayesian deep learning method is proposed for the model uncertainty evaluation of semantic segmentation. Semantic segmentation is one of the most important perception problems in understanding visual scene, which is critical for autonomous driving. This study to optimize Bayesian SegNet for uncertainty evaluation. This paper first simplifies the network structure of Bayesian SegNet by reducing the number of MC-Dropout layer and then introduces the pyramid pooling module to improve the performance of Bayesian SegNet. mIoU and mPAvPU are used as evaluation matrices to test the proposed method on the public Cityscapes dataset. The experimental results show that the proposed method improves the sampling effect of the Bayesian SegNet, shortens the sampling time, and improves the network performance.

Keywords Semantic Segmentation · Bayesian SegNet · Uncertainty Evaluation

1 Introduction

In recent years, with the fast development of deep learning in the field of artificial intelligence, autonomous driving has become possible. Autonomous driving is advantageous in reducing traffic accidents and alleviating traffic congestion, thus becoming popular in research and business fields [1–3]. Autonomous driving is mainly composed of three key systems: perception system, decision-making system, and control system, which are indispensable. For perception systems, autonomous vehicles perceive road traffic conditions through object detection, image classification and other related technologies. Among these technologies, semantic segmentation is a pixel-level classification of images, which is very important as it can be used to infer scene geometry and object relationships. However, in semantic segmentation, the content and appearance of the scene often change considerably, which causes safety issues for autonomous

driving. Safety is the first priority in autonomous driving, and ISO26262 [4] and ISO/PAS 21448 (also known as Safety of the Intended Functionality (SOTIF)) [5, 6] are the two main safety related standards used to address safety of electrical and electronic components. In particular, due to the development of deep learning, SOTIF is receiving increasing attention.

There are several limitations of SOTIF, such as complex and unstructured driving environments, and inherent security defects of deep learning models for autonomous driving, including the accuracy of sample label in the training set and the robustness within the operating range in an open environment. Deep learning has significant advantages in semantic segmentation, and lots of networks have been proposed such as fully convolutional networks [7], dilation networks [8], and SegNet [9]. Deep learning may introduce new challenges, such as non-transparency, error rate, training-based model, instability [10–13], which can affect the research on autonomous driving. Uncertainty is a natural part of the output of predictive system, especially the deep learning models. Knowing the confidence of the semantic segmentation output is important for decision-making system, several researchers are focusing on studying the uncertainty of perceptual deep learning. To better understand the feature learning and feature expression parts of the deep learning

Academic editor: Hong Wang

✉ Yang Zhao
yzhao@uestc.edu.cn

¹ School of Automation Engineering, University of Electronic Science and Technology of China, Chengdu, China

model, the interpretability of network was studied [14, 15]. Several Bayesian-based deep learning methods have been used to evaluate the uncertainty [16–19].

The main contribution of this study is simplifying the Bayesian SegNet and applying the pyramid pooling model to improve the performance of Bayesian SegNet. The remainder of this paper is organized as follows. In Sect. 2, an overview of the related work is presented. In Section 3, the details of the proposed method are discussed. Section 4 presents the experimental results. Finally, the study is summarized in Section 5.

2 Related Works

With the success of deep learning algorithms in the field of artificial intelligence, the research on the uncertainty of deep learning models has also been going on. Many researchers are attempting to evaluate the uncertainty in deep learning models by employing different methods. Ribeiro et al. [14] proposed LIME, which uses local linear approximation to model predictions to learn an interpretable model and use it to explain the predictions of any classifier. Their study also proposed a method of interpreting models by presenting representative individual predictions and their interpretation in a non-redundant manner. The task was built as a submodule optimization problem, and finally, the flexibility of the method was demonstrated by explaining different models of text and image classification. To better understand the feature learning and feature expression parts of the deep learning model, Dumitru et al. [15] studied the response of a single unit in the network, tested the output of different types of data, and obtained different expressions of the features using the activation function in the network. This was used to observe different model responses and to determine what type of data input can yield the maximum model output. Kendall et al. [20] presented a Bayesian deep learning framework to learn a mapping to aleatoric uncertainty from the input data, which are composed of input-dependent aleatoric and epistemic uncertainties. Moreover, they derived a framework for both regression and classification applications. Qi et al. [16] embedded a high-dimensional deep network layer nonlinearly into a low-dimensional interpretation space and used a few concepts extracted by the interpretation module to construct the original deep learning prediction. These concepts were then used to understand the advanced concepts that deep learning uses for decision making. They also embedded sparse reconstruction autoencoder (SRAE) [17] learning into the interpretation space and introduced some new indicators to quantitatively evaluate the performance of the interpretation. The experiments showed that the proposed method can better explain the mechanism CNN uses for prediction tasks. Hermann et al. [21] built

Fishyscapes based on the data from Cityscapes, which is a public benchmark for uncertainty estimation in the real-world task of semantic segmentation for urban driving.

In recent years, research on Bayesian deep learning (BDL) [18] has gained significant attention owing to its less calculation time and better interpretability. Rowan et al. [22] analyzed the impact of the uncertainty classification of autonomous driving perception systems on the decision-making systems, combined the advantages of a flexible deep learning architecture and the Bayesian method to propose a BDL framework, and finally reported that in the entire vehicle intelligence system, the subsystems at the bottom of the information flow need to be given greater attention. Kendall et al. [23] proposed a deep learning framework called Bayesian SegNet, which is based on probabilistic pixel-level semantic segmentation. This deep learning framework predicted the uncertainty of pixel-level labels and measurement models and generated a posterior distribution of pixel-like labels using Monte Carlo sampling with dropout (MC-Dropout) at the time of testing. The experiments conducted on the dataset showed that the uncertainty modeling improves the segmentation performance by 2–3%. To test the robustness of computer vision deep learning algorithms, Mukhoti et al. [24] proposed a framework for evaluating prediction uncertainty. Through the proposed framework, their study comprehensively compared the integrated method with MC-Dropout, and the results showed that the integrated method provides a more reliable uncertainty estimate, and the success of the integrated method is attributed to its ability to randomly initialize.

3 Proposed Pyramid Bayesian Method

This study aims to estimate the uncertainty of the semantic segmentation model using a new Bayesian deep learning (BDL) method with pyramid structure. Bayesian SegNet combines the original semantic segmentation network, SegNet [9], with the MC-Dropout [25] and obtains the semantic segmentation results and the uncertainty of the model simultaneously. Specifically, MC-Dropout is added to each phase layer of SegNet, which is applied to the normal iteration of the network according to the dropout layer during network training; however, it will not be closed during actual application and testing of the network. MC-Dropout method performs multiple forward propagation samplings on the network to obtain multiple semantic segmentation result graphs. Considering these semantic segmentations as a set, the average value and variance of this set are used as the semantic segmentation result and the uncertainty evaluation result of Bayesian SegNet, respectively. However, Bayesian SegNet framework reduces the speed in practical applications, and if the number of forward propagation samples

is reduced, the results of the algorithm become worse. To obtain accurate semantic segmentation results and uncertainty evaluation effect of the algorithm, it is necessary to increase the number of forward propagation samples, which can increase the operation time of the algorithm.

The information extracted from the shallow and the deep layers of a deep learning network is different. The shallow layers of the network often extract the corner and line information of images, and only some global information is extracted from the deep layer of the network. Therefore, MC-Dropout fusion for the deep layer of the network can extract network uncertainty more effectively and can reduce the time required for a single sampling. In addition, the pyramid pooling module is used to improve the sampling efficiency and reduce the total sampling times. In this section, BDL and MC-Dropout are first introduced and then the description of the proposed method is presented.

3.1 Bayesian Deep Learning and MC-Dropout

In the study of model uncertainty, the Bayesian method is often used to learn the probability distribution of weights in the model. This method changes each parameter in the neural network, that is, the weight and bias, from a certain value to a probability distribution. In conventional neural networks, with the model structure unchanged, the final results of the neural network training include the weight and bias of each layer of the network. However, in BDL, all weights and biases become probability distributions. Therefore, the network can only be trained and learned based on the method of multiple sampling parameters. Using Bayesian theory, the solution of the parameter weight probability distribution is transformed into the posterior probability of the parameter. There are generally three ways to estimate the posterior probability. The first method involves using Markov Chains Monte Carlo [26] sampling to approximate the probability distribution function, that is, Monte Carlo sampling. The second method involves using an indirect method to continuously approximate the true posterior probability distribution based on a simple probability distribution. The third method employs the most widely used method in deep learning, which belongs to the Monte Carlo sampling. This method integrates Bayesian statistical methods into neural networks using MC-Dropout to sample the network.

The Monte Carlo sampling method is a method of statistical simulation. When a probability distribution is known, Monte Carlo sampling is used to make the computer-generated sample data satisfy the probability distribution. Then, the appropriate sample data are used to explain and express the probability distribution. However, in deep learning models, it is very difficult to use conventional sampling methods to appropriately sample the probability distribution of weight parameters in the

network. During the training of the neural network, adding MC-Dropout to the appropriate network layer can considerably improve the generalization ability of the network and can also prevent the network from falling into the local best advantage in the training and eventually leading to overfitting [25, 27]. Gal et al. [28] considered the role of the MC-Dropout layer in neural networks as a Gaussian process in deep learning and proposed the MC-Dropout method to measure the uncertainty of models and algorithms. For a given input data, if the uncertainty obtained by the network is higher, the lower the confidence of the model on the output of these data, the more likely the output is incorrect.

3.2 Beyond Bayesian SegNet

The MC-Dropout for the deep structure of the network can extract network uncertainty more effectively and can reduce the time required for a single sampling. Based on the Bayesian SegNet network structure, the number of MC-Dropout layers is first reduced and the outermost MC-Dropout layer of the encoding–decoding structure is removed. The simplified Bayesian SegNet is shown in Fig. 1.

In semantic segmentation, to achieve accurate scene perception, most semantic segmentation networks integrate the edge information of the shallow network with the semantic information of the deep network as much as possible in the design of the network structure. A pyramid pooling module was proposed in PSPNet [29] to aggregate context information of different scales. PSPNet uses feature pooling layers of different scales in the deep layer of the network to convert feature maps to different sizes and then uses convolution kernels to extract information and finally combine them to improve the final segmentation result. Inspired by PSPNet, a pyramid sampling structure (PSS) is applied to improve the pooling layer. Pooling layers with different scales were used to resize feature maps to different sizes. Then, to improve the uncertainty estimation in semantic segmentation, MC-Dropout was added to feature maps of different sizes for the Monte Carlo sampling process. Finally, the sampled feature map was restored to its original size through a 1×1 convolution layer and was concatenated to ensure the size consistency of the network structure. The PSS is shown in Fig. 2, and the 1×1 , 2×2 and 3×3 pooling settings refer to PSPNet [29].

The final network framework is shown in Fig. 3. Based on the network in Fig. 1, we applied a PSS to improve the pooling layer. Finally, multiple forward propagation sampling methods were used to obtain the final semantic segmentation and uncertainty results.

Fig. 1 Simplified Bayesian SegNet

Fig. 2 Pyramid sampling structure

4 Test and Analysis

In this section, the dataset used to test the proposed method is first introduced. Subsequently, the criteria for semantic segmentation and uncertainty are presented. Finally, the experimental results and analysis are illustrated.

4.1 Cityscape Dataset

The Cityscapes dataset [32] is an urban environment semantic segmentation dataset for autonomous driving development, which is shown in Fig. 4. It is mainly a semantic segmentation image dataset for urban street scenes, which includes street scene data of 50 different cities mainly in Germany and neighboring countries. The dataset is divided into three types. The first type contains approximately 20,000 rough labeled images, the second type contains 5,000 frames of high-quality pixel-level semantic segmentation

Fig. 3 The proposed network

Fig. 4 The Cityscapes dataset

Fig. 5 Fine-labeled samples

labels and is commonly used in semantic segmentation tasks. Some fine-labeled samples are shown in Fig. 5. The third type consists of videos recorded from cars perspective, which is mainly used for the display and testing of some models. The number of training samples is 2,975, the number of tests is 1525, and the rest 500 samples are used for validation.

4.2 Uncertainty Evaluation Criteria

The output of the proposed network is the results of semantic segmentation and uncertainty. Therefore, the evaluation criteria are divided into two aspects for comparison. For the semantic segmentation task, the evaluation index used in this study is mean intersection over union (mIoU) [30, 31]. Semantic segmentation is essentially a classification task. As a classification task, the two classes are the ground truth and the predicted result. And the prediction result is divided into four cases: true positive (TP), false positive (FP), true negative (TN), the false negative (FN). mIoU is used to calculate the ratio of the intersection and union of the ground truth and the predicted result, which is

$$mIoU = \frac{TP}{(FN + FP + TP)} \quad (1)$$

mIoU is an important index to measure the accuracy of image segmentation results. IoU is calculated based on the divided categories. First, the IoU of each category is calculated and accumulated, then the average value is calculated

to obtain the final global evaluation, that is, mIoU. Equation (1) is equivalent to the following equation:

$$mIoU = \frac{1}{k+1} \sum_{i=0}^k \frac{p_{ii}}{\sum_{j=0}^k p_{ij} + \sum_{j=0}^k p_{ji} - p_{ii}} \quad (2)$$

where p_{ij} represents the number of pixels whose true value is i but is predicted to be j . Similarly, p_{ji} represents the number of pixels whose true value is j but is predicted to be i , and $k+1$ is the total number of categories including empty categories. p_{ii} represents the real quantity.

Model uncertainty evaluation is different from semantic segmentation task; however, both must use the same inputs. Evaluating whether an uncertain result is good is still a challenging problem, and so far only few related studies have been conducted on it. Mukhoti et al. proposed three evaluation indexes [24], that is, $p(\text{accurate/certain})$, $p(\text{uncertain/inaccurate})$, and patch accuracy vs. patch uncertainty (PAvPU). As PAvPU combines both the good cases of (accurate and certain) and (inaccurate and uncertain) patches into a single metric, PAvPU is selected as the evaluation index in the experiment.

The specific calculation process of PAvPU is as follows. For a Bayesian semantic segmentation model, the semantic segmentation and uncertainty results are obtained, along with the original label image serving as the three elements of the calculation. Then, based on the label and semantic segmentation results, a square sliding slider with a side length of w is used to synchronize the rolling sliding window. During each sliding process, based on the correctness threshold, it is estimated whether the accuracy rate in the window has been reached, and then, the segmentation correctness matrix is obtained. Similarly, the uncertainty image is sliding-windowed, and the uncertainty matrix is obtained based on the uncertainty threshold. Generally, the average value of uncertainty is used as the uncertainty threshold, that is, the mean of patch accuracy vs. patch uncertainty (mPAvPU). The mPAvPU formula is as follows.

$$mPAvPU = \frac{(n_{ac} + n_{iu})}{(n_{ac} + n_{au} + n_{ic} + n_{iu})} \quad (3)$$

where n_{ac} , n_{au} , n_{ic} , and n_{iu} represent the number of pixels in an image that are correct and determined, correct and uncertain, incorrect and determined, and incorrect and uncertain, respectively. Finally, the average value of all pictures in the statistical test set was used as the final evaluation criterion for mPAvPU, and a higher mPAvPU value represents a better uncertainty extraction result. In the experiment, mPAvPU is used as an evaluation index of uncertainty.

4.3 Experimental Analysis

In the experiment, a total of three groups of models with different structures were used for training, including Bayesian SegNet, simplified Bayesian SegNet (SBS) and SBS with PSS (SBS+PSS). The training and test datasets and corresponding training details of all models were kept consistent to ensure the effectiveness of experimental comparison. All models were built based on PyTorch.

The specific training details are as follows. The dataset used was fine-labeled semantic segmentation samples in the Cityscapes dataset, and a total of 5,000 frames of pictures were used as the original training dataset. In all the experiments, the images were first cropped to a size of 512×512 before being fed to the deep neural network. The random gradient descent method was used during training. The initial learning rate was set to 0.01, and the weight decay was set to 0.0005. The momentum was set to 0.9 and the batch size was set to 8. Moreover, the final loss function was cross entropy loss, and the number of training iterations was set to 60,000, divided into 120 epochs.

From Fig. 6, it can be observed that the SegNet model does not contain MC-Dropout layer; therefore, its mIoU value does not change with the increase in the number of samplings. However, for the other three types of BDL networks, the mIoU value continues to increase as the number of samplings increases, eventually reaching the upper limit of the accuracy of SegNet. This is because as the number of samplings increases, the corresponding semantic segmentation sampling set becomes larger, and its mean value continues to approach the true value of the accuracy of the model. The upper limit of the mIoU value of the semantic segmentation network fused with BDL is often higher than that of ordinary SegNet. For the SBS network, it can be observed that the change in accuracy is the same as that observed in Bayesian SegNet. However, the accuracy can sometimes be

lower than the original network's accuracy, which proves that the two outermost MC-Dropout layers in the original network have little effect on the accuracy. SBS+PSS can reach the upper limit of model accuracy in approximately 13 samplings, and Bayesian SegNet takes approximately 16 samplings. Moreover, the upper limit of the proposed model's mIoU value is higher than the Bayesian SegNet.

Figure 7 shows the relationship between the sampling times and mPAvPU values of three different models. As the ordinary SegNet network does not have the ability to extract uncertainty, it is not listed as a relevant test model in Fig. 7. It is known that the mPAvPU value of all BDL networks increases with the number of samples. A higher mPAvPU indicates a better uncertainty extraction effect, but it can barely reach 100%. When compared with the original Bayesian SegNet, the mPAvPU value of SBS network is slightly lower than the accuracy of the original network, which is acceptable. It proves that the two outermost MC-Dropout layers in the original network do not contribute in improving the ability to evaluate uncertainty. The proposed network can reach the upper limit of mPAvPU value in approximately 13 sampling, whereas Bayesian SegNet requires approximately 20 sampling, which proves that uncertain evaluation ability of SBS+PSS is better than the original network under the same sampling times.

As shown in Figs. 6 and 7, removing the shallow layer cannot significantly affect the semantic segmentation accuracy, but the PSS sampling of the shallow network in SegNet+PSP can result in longer sampling time. Therefore, this network structure is not used in the experiments. In addition, because the total time of the BDL network in actual application is closely related to its sampling times, the single sampling time of different models are also compared and results are shown in Table 1. It can be seen that the single sampling time of the SBS network is slightly shorter than the Bayesian SegNet, which is due to the deletion of two layers of MC-Dropout in the SBS network. The network

Fig. 6 Relationship between sampling times of different models and mIoU

Fig. 7 Relationship between sampling times of different models and mPAvPU

Table 1 Comparison of sampling time between different models

Methods	Single sampling time (ms)	Total sampling time (mPAvPU = 93%)	Total sampling time (mIoU = 72%)
Bayesian SegNet	20.4	408 ms (20 samples)	285.6 ms (14 samples)
SBS	20.2	404 ms (20 samples)	282.8 ms (14 samples)
SBS+PSS	24.3	340.2 ms (14 samples)	243 ms (10 samples)

Fig. 8 Experimental comparison in a simple scene **Fig. 9** Experimental comparison in a complex scene

introduces a pyramid sampling layer structure, which results in increased single sampling; however, the single-time uncertainty evaluation and semantic segmentation results were better than those obtained by Bayesian SegNet and therefore the mIoU was 72% (note that Bayesian SegNet and SBS reached 71.8% in Fig. 6. For a more convenient comparison, it is approximated to 72% in Table 1.). The proposed method only took 243 ms, whereas Bayesian SegNet and SBS took 285.6 ms and 282.8 ms, respectively. When the mPAvPU value was 93%, the proposed network only needed to sample 14 times, whereas Bayesian SegNet needed to sample 20 times, which shows that the proposed network needs less time to sample.

In summary, the proposed network improved the result of single sampling with a slight sacrifice of single sampling time, reduced the number of samplings, and ultimately shortened the total sampling time. The model generates simple and complex scenes (shown in Figs. 8 and 9, respectively). Simple scenes and complex scenes are distinguished by evaluating the amount of semantic information in the picture, such as the number of types of objects. Simple scenes generally contain only fewer objects, such as sky, straight roads, trees, buildings, whereas complex scenes usually contain richer semantic information, such as clouds in the

sky, unstructured roads, traffic signs, crosswalks, pedestrians, and vehicles. In the corresponding uncertainty result images, black indicates low uncertainty and bright areas indicate high uncertainty. The uncertainty in scenes often appears at the boundaries of different categories, and the uncertainty within the categories is very low, which represents the normal situation that conforms to the model classification. In the case of 10 samples, the proposed network had fewer errors in the semantic segmentation results than Bayesian SegNet. Moreover, the proposed network was more efficient in evaluating the uncertainty than Bayesian SegNet.

5 Conclusions

This study proposes a novel pyramid BDL network for the model uncertainty evaluation of semantic segmentation. To better evaluate the uncertainty of semantic segmentation network model, the number of MC-Dropout layer is reduced to simplify Bayesian SegNet and the pyramid pooling model is used to improve the performance of Bayesian SegNet. mIoU and mPAvPU serve as reference evaluations on the public Cityscapes dataset, and the results show that the proposed method improves the sampling effect of the

Bayesian SegNet, shortens the sampling time, and improves the network performance.

However, as MC-Dropout requires multiple stochastic runs during test time, its high computational cost prevents uncertainty estimation with the MC-Dropout and is still infeasible for real-time critical systems. In the future work, MC-Dropout will be optimized to improve the efficiency and real-time performance of uncertainty estimation.

Acknowledgements This work was supported by the National Natural Science Foundation of China (U1964203) and the National Key R&D Program Project of China (2017YFB0102603).

Declarations

Conflict of Interest We declare that we do not have any commercial or associative interest that represents a conflict of interest in connection with the work submitted.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. WHO: Global status report on road safety 2015. World Health Organization, Violence and Injury. Prevention (2015)
2. NSC: Motor vehicle fatality estimates. Technical report, National Safety Council, Statistics Department (2016)
3. Wang, H., Huang, Y., Khajepour, A., et al.: Ethical decision-making platform in autonomous vehicles with lexicographic optimization based model predictive controller. *IEEE Trans. Vehi. Technol.* **69**(8), 8164–8175 (2020)
4. Schildbach G.: On the application of ISO 26262 in control design for automated vehicles. *arXiv preprint. arXiv:1804.04349* (2018). Accessed 12 Apr 2018
5. Wendorff, W.: Quantitative SOTIF analysis for highly automated driving systems. In: *Safetrone Conference*, Stuttgart (2017)
6. Pimentel J.: Automated vehicles: safety of the intended functionality (SOTIF). SAE (2020)
7. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **39**(4), 640–651 (2015)
8. Yu, F., Koltun, V.: Multi-scale context aggregation by dilated convolutions. *arXiv:1511.07122* (2016). Accessed 30 Apr 2016
9. Badrinarayanan, V., Kendall, A., Cipolla, R.: SegNet: a deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **39**(12), 2481–2495 (2017)
10. Salay R., Queiroz R., Czarnecki K.: An analysis of ISO 26262: Using machine learning safely in automotive software. *arXiv:1709.02435v1* (2017). Accessed 7 Sep 2017
11. Abdar M., Pourpanah F., Hussain S., et al.: A review of uncertainty quantification in deep learning: techniques, applications and challenges. *arXiv:2011.06225* (2020). Accessed 17 Nov 2020
12. Nascimento, A.M., Vismari, L.F., Molina, C.B.S.T., et al.: A systematic literature review about the impact of artificial intelligence on autonomous vehicle safety. *IEEE Trans. Intell. Transp. Syst.* **21**(12), 4928–4946 (2020)
13. Feng D., Harakeh A., Waslander S., et al.: A review and comparative study on probabilistic object detection in autonomous driving. *arXiv:2011.10671* (2020). Accessed 20 Nov 2020
14. Ribeiro M.T., Singh S., Guestrin C.: Why should i trust you? explaining the predictions of any classifier. *arXiv:1602.04938v2* (2016). Accessed 9 Aug 2016
15. Erhan, D., Bengio, Y., Courville, A., et al.: Visualizing higher-layer features of a deep network. *Univ. Mont.* **134**(3), 1 (2009)
16. Qi, Z., Li, F.: Learning explainable embeddings for deep networks. In: *Thirty-first Conference on Neural Information Processing Systems Workshop on Interpreting, Explaining and Visualizing Deep Learning*, Long Beach, USA, 4–9 Dec 2017
17. Ng, A.: Sparse autoencoder. *CS294A Lecture notes* **72**(2011), 1–19 (2011)
18. Wang, H., Yeung, D.Y.: Towards bayesian deep learning: a framework and some existing methods. *IEEE Trans. Knowl. Data Eng.* **28**(12), 3395–3408 (2016)
19. Feng, D., Rosen, B.U.L., Dietmayer, K.: Towards safe autonomous driving: capture uncertainty in the deep neural network for lidar 3D vehicle detection. In: *IEEE International Conference on Intelligent Transportation Systems*, Hawaii, USA, 4–7 Nov 2018
20. Kendall, A., Gal, Y.: What uncertainties do we need in Bayesian deep learning for computer vision? *Thirty-first Conference on Neural Information Processing Systems*, pp. 4–9. Long Beach, USA 4–9 Dec 2017
21. Blum, H., Sarlin, P.E., Nieto, J., et al.: Fishyscapes: a benchmark for safe semantic segmentation in autonomous driving. In: *International Conference on Computer Vision Workshop*, Seoul, Korea, 20–26 Oct 2019
22. McAllister, R., Gal, Y., Kendall, A., et al.: Concrete problems for autonomous vehicle safety: advantages of bayesian deep learning. In: *International Joint Conferences on Artificial Intelligence*, Melbourne, Australia, 19–25 Aug 2017
23. Kendall, A., Badrinarayanan, V., Cipolla, R.: Bayesian segnet: model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. *arXiv:1511.02680* (2015). Accessed 9 Nov 2015
24. Mukhoti J., Gal Y.: Evaluating bayesian deep learning methods for semantic segmentation. *arXiv:1811.12709* (2018). Accessed 30 Nov 2018
25. Srivastava, N., Hinton, G., Krizhevsky, A., et al.: Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **15**(1), 1929–1958 (2014)
26. Gustafsson, F.K., Danelljan, M., Schn, T.B.: Evaluating scalable bayesian deep learning methods for robust computer vision. *arXiv:1906.01620* (2019). Accessed 4 Jun 2019
27. Krizhevsky, A., Sutskever, I., Hinton, G.: ImageNet classification with deep convolutional neural networks. In: *Twenty-sixth Conference on Neural Information Processing Systems*, Nevada, USA, 3–8 Dec 2012
28. Gal, Y., Ghahramani, Z.: Dropout as a bayesian approximation: representing model uncertainty in deep learning. In: *International Conference on Machine Learning*, New York City, USA, 26 June–1 July 2016

29. Zhao, H., Shi, J., Qi, X., et al.: Pyramid scene parsing network. In: IEEE Conference on Computer Vision and Pattern Recognition, Hawaii, USA, 21–26 July 2017
30. Everingham, M., Eslami, S.M.A., Van Gool, L., et al.: The pascal visual object classes challenge: a retrospective. *Int. J. Comput. Vis.* **111**(1), 98–136 (2015)
31. Berman, M., Rannen Triki, A., Blaschko, M.B.: The lovsz-softmax loss: a tractable surrogate for the optimization of the intersection-over-union measure in neural networks. In: IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, USA, 18–22 June 2018
32. Cordts, M., Omran, M., Ramos, S., et al.: The cityscapes dataset for semantic urban scene understanding. In: IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 26 June–1 July 2016

Yang Zhao received the Ph.D. degree in Control Science and Engineering from University of Electronic Science and Technology of China (UESTC), in July 2018. He joined School of Automation Engineering, UESTC, as an assistant researcher fellow, in 2018. He was a joint training doctoral student with University of Washington, from 2015 to 2016. His current research interests include intelligent connected vehicles and human-machine interaction.

Wei Tian received the M.S. degree in Pattern Recognition and Intelligent Systems from University of Electronic Science and Technology of China (UESTC), in July 2019. His current research interests include computer vision and deep learning.

Hong Cheng received the Ph.D. degree in Pattern Recognition and Intelligent Systems from Xi'an Jiaotong University, in 2003, where he has been an associate professor, since 2005. He held a postdoctoral position with the Computer Science School, Carnegie Mellon University, USA, from 2006 to 2009. Since July 2000, he has been with Xi'an Jiaotong University, where he was a Team Leader of the Intelligent Vehicle Group, Institute of Artificial Intelligence and Robotics, until 2006. He is a full professor with the School of Automation and the Director of Center for Robotics, UESTC, where he is currently the founding director of the Machine Intelligence Institute. His current research interests include computer vision and machine learning, robotics, human computer interaction, and multimedia signal processing.