This article has been accepted for publication in IEEE Transactions on Intelligent Vehicles. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIV.2022.3204714

Curiosity-Driven Attention for Anomaly Road Obstacles Segmentation in Autonomous Driving

Xiangxuan Ren, Min Li, Zhenhua Li, Wentao Wu, Lin Bai, and Weidong Zhang, *Senior Member, IEEE*

***Abstract*—The inability of semantic segmentation methods to detect anomaly road obstacles not pre-defined in the datasets significantly hinders the safety-critical application in autonomous driving. The excessively complex anomaly detection approaches cannot accommodate the constraints on the inference time of intelligent vehicles. Inspired by the fact that humans have a natural instinct to be curious about unknown objects in a new environment, we propose a novel curiosity-driven attention mechanism (CuDAM) for anomaly road obstacles segmentation. CuDAM adopts the attention map as a new uncertainty judging criterion and utilizes it to improve the efficiency of the model. Specifically, CuDAM is composed of three parts: 1) an attention module for generating an attention map; 2) a reward mechanism for encouraging the network to focus its attention on uncertain regions; 3) an attention loss function for widening the distance between the attention values of deterministic and uncertain pixels. Different from previous approaches, CuDAM can improve both anomaly detection and semantic segmentation performance without complex operations and training, which makes it widely applicable to existing semantic segmentation models. The result of qualitative and quantitative experiments shows that such a straightforward approach achieves consistent significant improvements in anomaly detection performances with the various uncertainty estimation methods, demonstrating the broad applicability of CuDAM.**

***Index Terms*—Autonomous driving, anomaly detection, CuDAM, attention module, semantic segmentation.**

I. Introduction

THE last decades marked tremendous progress in autonomous driving [1]–[8]. Self-driving vehicles are urgently demanded in life scenarios or dangerous environments which are unreachable to humans, and safety is a crucial concern for its application. Autonomous driving platforms build on accurate visual perception systems [9]–[11], in which semantic segmentation is an essential technology for pixel-wise classification of camera images. Recent studies in semantic segmentation focus on how to improve the precision of segmentation performance. However, the highly accurate pixel-level classification of objects is based on the premise of a strongly supervised deep learning approach by training models using large, fully annotated datasets. The segmentation models can only classify pre-defined categories in the datasets, which implies an overly idealistic assumption that all possible objects are included in the training set. Unfortunately, the real world is open and unexpected things may happen at any time. Faced with an unknown abnormal target (e.g., a cat suddenly jumping onto the road in Fig. 1), models may fail to recognize it and overconfidently classifies it into a pre-defined class. This scenario creates a severe safety hazard and greatly limits the application of deep learning algorithms in autonomous driving. Also, it is not practical to collect a dataset containing all types of obstacles that may appear on the road. Therefore, for a perception network, it is essential to train it to have the ability to identify anomalous obstacles on the road.

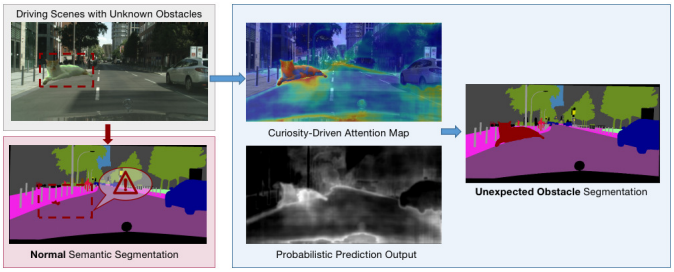

Fig. 1. Anomaly scenarios overview. The semantic segmentation model overconfidently classifies the anomaly obstacles as one of the pre-defined classes, which causes serious security risks (as pointed out by the red arrows). We propose a CuDAM method that makes the model focus on the abnormal target and obtains the final prediction of segmentation labels with unexpected road obstacles identified.

The problem of enabling models to detect anomaly road obstacles has been tackled in several works of [12]–[16]. Based on the intuition that anomalous targets tend to have low prediction probabilities, the uncertainty estimation approaches are one class of methods to obtain anomaly scores by designing different functions to calculate the uncertainty probabilities. However, since the models are often overconfident about the anomalous objects, the detection results are noisy and inaccurate. Another major class of methods accomplishes anomalous obstacle detection by adding additional training tasks. Some approaches train the model by utilizing external out-of-distribution (OoD) datasets as samples in this class. In contrast, others leverage feature reconstruction methods to manually design or learn features of unknown classes to distinguish anomalous classes. The generated models are also adopted by some methods to re-synthesize the input images.

This paper is partly supported by Shanghai Science and Technology program (22015810300; 19510745200), Hainan Province Science and Technology Special Fund (ZDYF2021GXJS041), and the National Natural Science Foundation of China (U2141234), in part by the Hainan Special PhD Scientific Research Foundation of Sanya Yazhou Bay Science and Technology City under Grant HSPHDSRF-2022-01-005 and Grant HSPHDSRF-2022-01-007. (*Corresponding Author: Weidong Zhang.*)

X. Ren, M. Li, Z. Li, W. Wu, and W. Zhang are with the Department of Automation, Shanghai Jiao Tong University, Shanghai 200240, China, L. Bai is with Guangzhou Gosuncn Robot Co. Ltd., Guangzhou 510000, China, and W. Zhang is also with the School of Information and Communication Engineering, Hainan University, Haikou 570228, Hainan, China (e-mail: bunny_renxiangxuan@sjtu.edu.cn; li_min@sjtu.edu.cn; lizhenhuagd@163.com, lizhenhuagd@sjtu.edu.cn; wentao-wu@sjtu.edu.cn; bailin@gusuncn.com; wdzhang@sjtu.edu.cn).

Although such works have been validated to be effective, they necessitate a lengthy inference time or add a non-trivial amount of labor intensity. Meanwhile, retraining may decrease the semantic segmentation performance of the original network. Consequently, an exciting question naturally arises: Can we design an anomaly obstacles segmentation approach by exploring a balanced solution among the two categories of methods mentioned above? This method improves the performance of the uncertain method without significantly increasing the computational amount and training difficulty, and without affecting the semantic segmentation accuracy.

We are inspired by the fact that humans are naturally curious about objects they are not familiar with [17]. This curiosity drives us to pay more attention to unknown objects in new scenarios to improve learning efficiency. This motivates us to use attention as a new measure of whether an object is unexpected or not. As shown in Fig. 1, the curiosity-driven attention can both help the model identify anomalous targets and improve the training efficiency of the neural network by enabling the network to focus on learning uncertain regions and less on already proficient categories. Implementing the above approach involves two practical problems: a) how to train the network to have this curiosity attention to focus on anomalous obstacles. b) how to use the curiosity-driven attention graph to identify anomalous obstacles.

In this paper, we propose a CuDAM approach for anomaly road obstacles segmentation within autonomous driving. First, an attention map of the same size as the intermediate input features is generated based on the intermediate layer features. Then, the softmax function is taken to recalculate the attention values, which ensures that the sum of the attention map is a certain number. The setting forces the model to allocate attention rationally with limited attention. Based on this, we propose a reward mechanism similar to reinforcement learning to give appropriate rewards based on the attention map, which encourages the model to focus on uncertain regions. Finally, an attention loss function is proposed to further enlarge the distance of the attention value between certain and uncertain pixels. The above approach allows us to train a semantic segmentation network incorporating the CuDAM. Anomalous obstacle segmentation can be achieved by integrating the attention graph with the uncertainty score.

CuDAM is a plug-and-play method that can be widely applied in semantic segmentation networks to improve the performance of anomalous obstacle detection while enhancing semantic segmentation capabilities. Quantitative and qualitative experiments are conducted to show the contribution of our approach. To analyze the effects of attention mechanisms from an interpretable perspective, we then visualize the attention model and compare the effects of different experimental settings on attention.

The main contributions of our work are as follows:

- We present a CuDAM method for anomaly road obstacles segmentation, which is a novel trial to utilize an attention mechanism to tackle the anomaly detection problem.
- The curiosity-driven attention is a new form of attention mechanism closer to the laws of human cognition, which can be used to enhance anomaly detection performance based on various uncertainty methods.
- The proposed method improves the performance of both pixel-wise anomaly detection and semantic segmentation without significantly increasing the number of parameters and complex training processes.

II. Related Works

Recent works [3]–[8] on semantic segmentation [4], object detection [3], small object detection [6], [7], lane line detection, traffic light recognition, and more have led to breakthroughs in improving autonomous driving perception performance. Despite the advances, anomaly detection is still a safety-critical task. The task of anomaly detection is to address security concerns in real-world scenarios where the system is confronted with unknown targets that do not exist or rarely occur in the dataset. A large part of anomaly detection work is focused on the image level, while pixel-wise anomaly segmentation as an intensive prediction task is more challenging. In this section, we review the approaches which could be used for pixel-wise anomaly segmentation within autonomous driving. We divide these methods into two broad categories: the method based on uncertainty estimation and the method based on introducing additional training tasks.

A. Anomaly Segmentation via Uncertainty Estimation

The most straightforward method to detect anomalies is based on uncertainty estimation approaches which are driven by searching for an effective uncertainty calculation function. The uncertainty can be interpreted as a pixel-wise anomaly score to detect unexpected obstacles on the road. A higher uncertainty score means a higher abnormal score. As one of the initial works, [18] derives the uncertainty score by using Bayesian neural networks with a Monte Carlo dropout estimate. Several follow-up studies in [19], [20] tackle the problem utilizing the Bayesian method. However, Bayesian segmentation networks are slow in inference due to their multiple forward passes in the network, with Monte Carlo dropouts in each frame. A more natural and simple intuition is that a normal object has a higher maximum softmax probability (MSP) than an (OoD) sample [21]. Alternatively, max logits have been shown to be a more valid value for assessing uncertainty than MSP in [22]. The work of [23] proposes a standardized calculation method to obtain class-conditioned standardized max logits (SML) to replace max logits to evaluate uncertainty. While these methods are effective baselines for image-level anomaly detection, they tend to misclassify object boundaries as anomalies. SML suppresses class boundaries and applies a dilated smoothing to consider local semantics, but the challenge of false positives at the boundary still exists. Without fine-tuning with additional outlier data, it is difficult for the uncertainty-based anomaly segmentation methods to obtain accurate OoD detection results due to the misclassification caused by over-confidence and boundary false positive problems.

B. Anomaly Segmentation via Introducing Additional Training Tasks

Several studies utilize additional training tasks for anomaly detection. These detection approaches can be divided into three concepts: feature reconstruction, auxiliary dataset, and image re-synthesis. Anomaly segmentation via feature reconstruction uses handcrafted or learned features to determine a class label [13], [16], [24]–[28]. The works of [16], [26], [27] suggest that anomalies can be segmented by reconstructing the normality of input and considering any kind of deviation from its anomalous. Although those approaches achieve excellent performance at the object level, they are challenged by the dependence on an accurate pixel-wise segmentation prediction, by the complexity of reconstruction models, and also by the low quality of the reconstructed features. Approaches based on auxiliary datasets utilize external datasets as samples of unexpected objects to improve the anomaly detection performance in [29], [30]. The works of [31], [32] modify the segmentation network as a muti-task model and train a network to differentiate inliers against OoD samples. However, the occurrence of corner cases is extremely accidental and random, it is impractical to collect all possible anomalies to train the model, which weakens the portability of the algorithm. Instead of using existing data directly, image re-synthesis-based methods are proposed in [15], [33], [34], which leverage autoencoders to synthesize anomalous samples. The works of [35], [36] assume that the reconstructed image region can better retain the features of the known class region compared to the OoD samples, thereby distinguishing the unknown obstacles. More recent methods use a generative adversarial network (GAN) to re-synthesize the input image from the predicted semantic map in [12], [15]. However, such approaches require a considerable amount of labor intensity or necessitate a lengthy inference time that is critical in semantic segmentation. While the additional training allows the model to achieve better results in anomaly detection than the uncertainty-based method, it also leads to lower performance in the original semantic segmentation task due to the challenge of balancing the training effects of different complex tasks.

III. Methodology

This section presents our CuDAM method for anomalous road obstacles segmentation. We first present our motivation in subsection A, then introduce the overall framework of CuDAM in subsection B, the attention module in subsection C, and the reward mechanism in subsection D. Finally, The training and inference procedures are described in subsection E.

A. Motivation

Humans have a natural instinct to identify unknown object instances in the environment quickly. This curiosity-driven attention can help humans learn new things more efficiently. This motivates us to propose a new CuDAM method. Our overall thinking is to train the model to focus on its uncertain regions so that it can have human-like curiosity. By using this attention map and combining it with the output of the network, the vehicles can naturally learn to recognize unexpected obstacles on the road. A higher attention span means that this area is more likely to be abnormal.

To implement this idea, a reward mechanism is proposed which is similar to reinforcement learning. Since high attention value areas deserve more helpful information, the regions with high attention values are rewarded with ground truth (GT) information to motivate the model to learn the decision boundary of the attention map. This inspiration is also similar to the teacher-student mechanism of the life scenario, where the student learns to admit their lack of confidence in identifying the category of a certain position, rather than just giving an over-confidence wrong answer. Then the teacher takes responsibility for answering the uncertain question of the student. Following the above simple strategies, the curiosity-driven attention approach is proposed to identify unexpected road obstacles in complex driving scenes.

B. Overall Architecture

Our method overview is illustrated in Fig. 2. Semantic segmentation models can generally be summarised as an encoder-decoder structure. The input image is initially transformed into high-dimensional features by the encoder. Then, given this intermediate feature as input, CuDAM first infers a 2D attention map. In particular, attention should not be infinite. A constant sum of attention values will force the entire pixels of the attention map to compete with each other in order to obtain the maximum benefit. This avoids the loophole that the model negatively sets all attention values to be large. Afterward, we select the multi-level multi-scale features generated by encoders fused with the attention map and input them into the decoder network to obtain the prediction output. After obtaining the attention graph and the prediction output, we adopt a reward mechanism that combines the correct segmentation answer with the output according to the attention value. This motivates the model to learn to assign high attention values to regions with high uncertainty. To widen the gap between the attention values of certain and uncertain pixels, we add a penalty to the loss function, i.e., the CuDAM loss. Finally, we combine the prediction output of the network with CuDAM's attention map to obtain the final ensemble prediction.

C. Attention Module

The following describes the details of the attention module. Given an intermediate feature map $\boldsymbol{F_a} \in \mathbb{R}^{C' \times H' \times W'}$ as input, CuDAM infers an attention map $\boldsymbol{M_{att}} \in \mathbb{R}^{1 \times H' \times W'}$. The higher the value on the graph $\boldsymbol{M_{att}}$, the less confident the model is in predicting the pixel class. We first compress the feature $\boldsymbol{F_a}$ using average-pooling and max-pooling operations along the channel axis to generate an efficient feature description similar to the CBAM [37]. Then, we obtain an attention map of the same size as the input feature $\boldsymbol{F_a}$ by concatenating them and feeding the feature into a standard convolution layer. To be specific, the softmax function is used to limit the sum of the attention value to a fixed value, producing our curiosity-driven attention map $\boldsymbol{M_{att}}$. The arrangement of using softmax instead of the sigmoid function is to make the

This article has been accepted for publication in IEEE Transactions on Intelligent Vehicles. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIV.2022.3204714

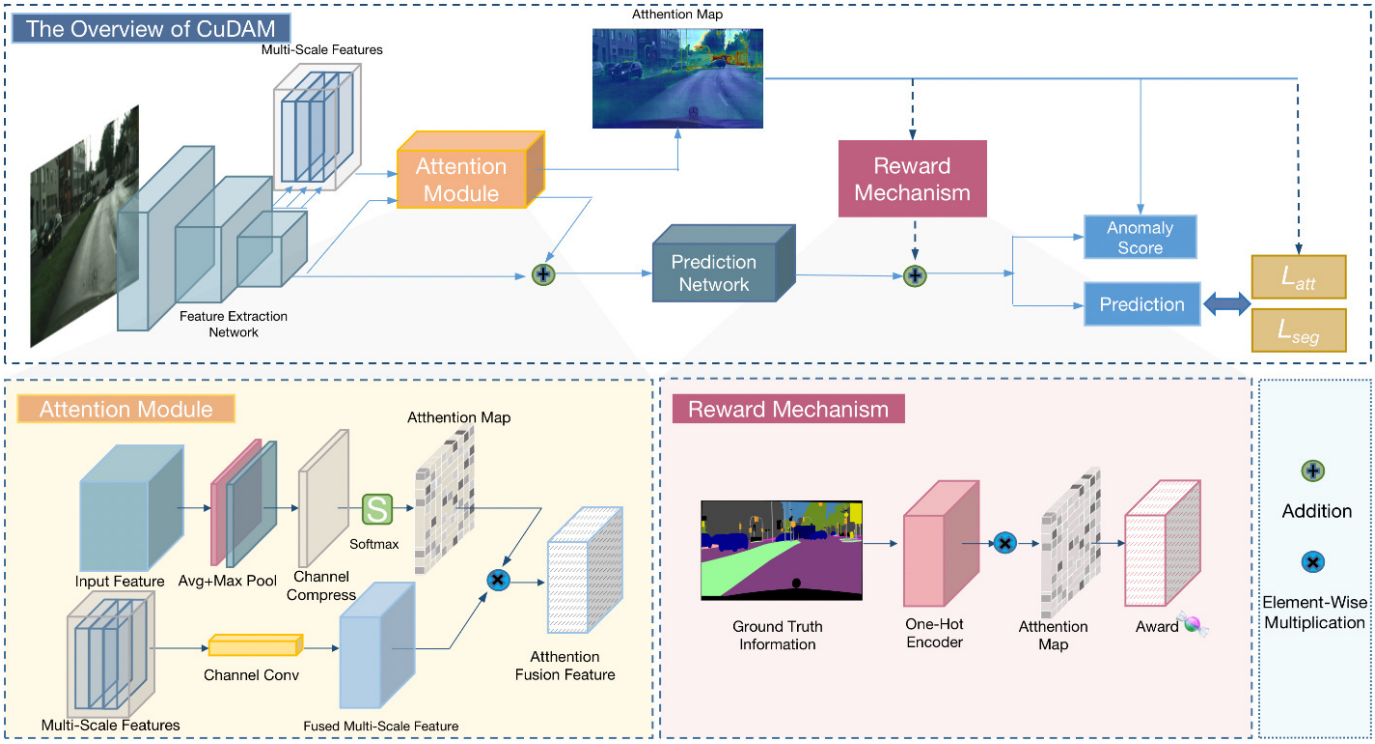

Fig. 2. The overview framework of the proposed CuDAM approach. The dotted line indicates that the process will only be used during the training phase. The attention module is used to generate a 2D attention map of the same size as the input feature to filter multi-scale information for network prediction. The reward mechanism encourages the network to focus on unknown targets by rewarding the network with different values of GT information according to the attention graph.

different pixel points compete for limited attention and thus achieve an optimal fit. In short, the attention map is computed as:

$$\boldsymbol{M_{att}}(\boldsymbol{F_a}) = \delta(f([Avgpool(\boldsymbol{F_a}); Maxpool(\boldsymbol{F_a})])) \quad (1)$$

where δ denotes the softmax function and f represent a standard convolution operation.

To obtain more detailed features for model inference in the attention-high region, we fuse multi-scale feature maps $\boldsymbol{F_{low}}$ and $\boldsymbol{F_{mid}}$, which are sequentially selected from two different layers of the encoder. Concatenation and summation are two common feature fusion methods, and our experimental result shows that the concatenation is better than the summation manner. We use a convolutional layer with a kernel size of 1×1 and an upsampling layer to convert the feature to be the same size as $\boldsymbol{F_a}$. The fusion multi-scale feature is computed as:

$$\boldsymbol{F_{sc}} = \eta(Concat(\boldsymbol{F_{low}}, \boldsymbol{F_{mid}}, \boldsymbol{F_a})) \quad (2)$$

where η denotes the feature processing operations using convolution and upsampling. The overall attention process can be summarized as:

$$\boldsymbol{F} = \alpha \boldsymbol{M_{att}}(\boldsymbol{F_a}) \otimes (\boldsymbol{F_a} + \boldsymbol{F_{sc}}) + \boldsymbol{F_a} \quad (3)$$

where $\otimes$ denotes element-wise multiplication and α is set as a learnable parameter that is gradually modified during training.

D. Reward Mechanism

Let $\boldsymbol{X} \in \mathbb{R}^{3\times H\times W}$ denote the input image, where H and W are the height and weight of the input image, respectively. C denotes the number of pre-defined categories satisfying $c \in \{1, \ldots, C\}$. Prior to normalization, the segmentation model generates a predicted probability $p_{c,h,w}$ for each class at each location h, w. Then, the entire logit output can be denoted as $\boldsymbol{P_a} \in \mathbb{R}^{C\times H\times W}$. The correct classification answer for the input image is recorded in ground truth $\boldsymbol{G} \in \mathbb{Z}^{H\times W}$, and $g_{h,w}$ in $\boldsymbol{G}$ indicates the class number c corresponding to the position h, w. The GT information only contains normal classes and not OoD samples.

We generate an award map $\boldsymbol{G_{award}} \in \mathbb{Z}^{C\times H\times W}$ by expanding the ground truth value $g_{h,w}$ at each location from a one-dimensional class number to a three-dimensional vector, which has a value of 1 for the correct category of positions c, h, w and 0 for the rest. Such a one-hot encoding method is adopted to convert the reward matrix to the same size as the predicted output.

We first up-sampling the attention map to the same size as the input image. This attention map is denoted as $\boldsymbol{M_{att}^{up}} \in \mathbb{R}^{1\times H\times W}$. Next, the reward feature is fed into the model to help the network make predictions:

$$\boldsymbol{P} = \tau \boldsymbol{M_{att}^{up}}(\boldsymbol{F_a}) \otimes \boldsymbol{G_{award}} + \boldsymbol{P_a} \quad (4)$$

where τ is reward hyper-parameter that controls the proportion of the reward. The intuition behind Eq. (4) is that high attention value areas deserve more helpful information. A higher attention value for each location h, w means that the model can be rewarded more, increasing the probability of prediction for the correct category by more. The GT reward value increases the predicted probability value of the model

This article has been accepted for publication in IEEE Transactions on Intelligent Vehicles. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIV.2022.3204714

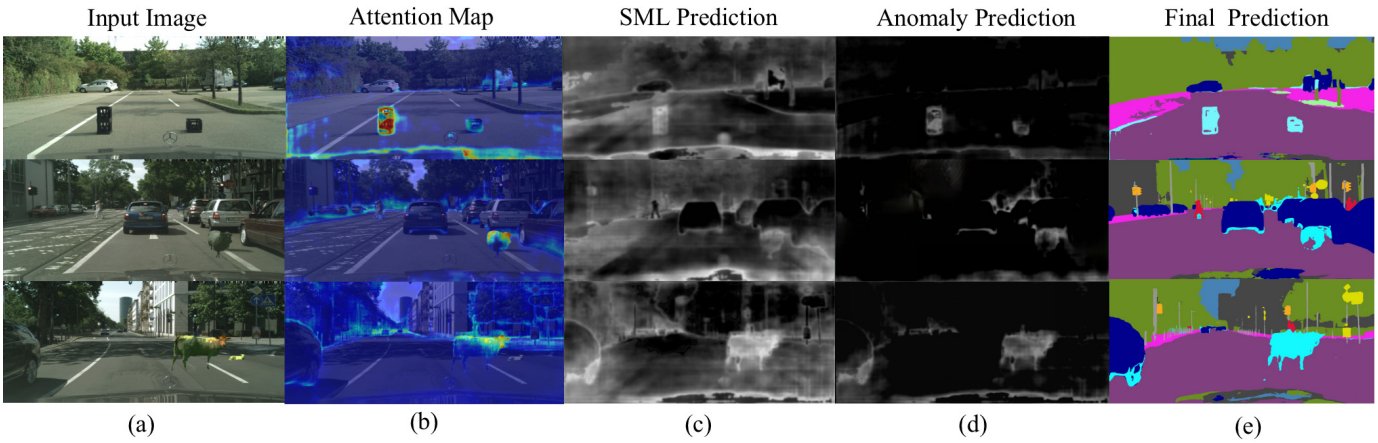

Fig. 3. Results of the CuDAM approach on identifying anomaly road obstacles within autonomous driving. Given the input images shown in (a), CuDAM can compute a curiosity-driven attention map by designing the reward mechanism and L_{att} as shown in (b). Next, the anomaly score is obtained based on the probability output predicted by the model, which is shown in (c). Then, as shown in (d), CuDAM ensembles the attention map and SML prediction using a weighted average to output the anomaly prediction. Finally, the segmentation results containing anomalous obstacles are obtained by integrating the anomaly prediction results with the semantic segmentation results as shown in (e).

output in the correct category dimension, which then reduces the loss function of the classification problem. In addition, to prevent the addition of the reward mechanism from making the model lazily over-reliant on rewards and leading to lower performance in semantic segmentation, we randomly choose whether to give a reward to the model in each iteration with a certain probability. This forces the model to optimize in both exam and study scenarios constantly. The model is not allowed to receive a reward during the exam scenario. In contrast, the reward mechanism is used to motivate the model to recognize its shortcomings during the study scenario.

E. Training and Inference

The final logit output $\boldsymbol{P}$ can be obtained from the previous reward method. Then, the max logit score $\boldsymbol{S} \in \mathbb{R}^{H\times W}$ and prediction $\boldsymbol{Y} \in \mathbb{R}^{H\times W}$ are defined as:

$$\boldsymbol{S_{h,w}} = \max_{c} \boldsymbol{P_{c,h,w}} \tag{5}$$

$$\boldsymbol{Y_{h,w}} = \arg\max_{c} \boldsymbol{P_{c,h,w}} \tag{6}$$

As proposed in uncertainty estimation methods in [21]–[23], different uncertainty estimation functions $g(\boldsymbol{S})$ can be selected to calculate the final anomaly score denoted as $\boldsymbol{K} \in \mathbb{R}^{H\times W}$.

In our experiment, we find that the margin between the attention values of the anomalous and normal targets is small using only the reward mechanism. To increase the gap between attention values, we added an attention loss function term to the original segmentation loss. The attention map $\boldsymbol{M^{up}_{att}}$ can be obtained from the previous reward method. Then, the mean of all predicted correct position attention value ψ is calculated. The correct position is the location that the model can predict correctly itself without using the GT reward, which is calculated from the $\boldsymbol{P_a}$, the opposite is the incorrect position. The intuition behind this is that the model should have a lower attention value on the position that it can easily predict correctly on its own. Next, the CuDAM attention loss is calculated as follows:

$$L_{\text{att}} = \begin{cases} \sum_{h,w} m_{h,w} e^{k_{h,w}}, & m_{h,w} < \beta \cdot \psi \\ 0, & m_{h,w} \geq \beta \cdot \psi \end{cases} \tag{7}$$

where $m_{h,w}$ denotes the attention value corresponding to the position h, w that is predicted incorrectly in $\boldsymbol{M^{up}_{att}}$ and β is an adjustable penalty hyper-parameter, which balances the gap between attention values of each position. The smaller β is, the greater the difference between the attention value of abnormal and normal pixels. Assuming that the loss function of the original semantic segmentation network is L_{seg}, then the total loss function is defined as:

$$L_{\text{total}} = L_{\text{seg}} + \lambda L_{\text{att}} \tag{8}$$

where λ is the weight parameter of L_{att}. Since the segmentation ability of the network is poor at the beginning of the training, we want the network to focus on improving the semantic segmentation accuracy. As the number of training sessions increases, we expect the network to focus on improving the semantic segmentation accuracy. As the number of training epochs increases, we expect the model to focus on learning how to improve the accuracy of anomaly detection by reasonably allocating attention. Therefore, the attention weight should be gradually increased with the training accuracy. Based on this idea, we design a self-regulation method of λ. We count the number of all correctly predicted pixels a_{correct}, and the number of all pixels involved in the loss function calculation a_{total}, λ is calculated as :

$$\lambda = \frac{a_{\text{correct}}}{a_{\text{total}}} \tag{9}$$

Until now, a curiosity-driven attention map $\boldsymbol{M_{att}}$ and a probability prediction output $\boldsymbol{K}$ can be obtained. The visualization results of the attention map are shown in Fig. 3(b). $\boldsymbol{K}$ can be calculated using various uncertainty estimation functions and we choose SML [23] to generate anomaly scores and display the visualization results in Fig. 3(c). Finally, we

ensemble the max logit score and the attention map using a weighted summation:

$$\boldsymbol{P}_{\mathbf{ano}} = \boldsymbol{K} + \sigma \boldsymbol{M}_{\mathbf{att}}^{\mathbf{up}} \tag{10}$$

where σ denotes anomaly score weighting hyper-parameter. With this simple ensemble, CuDAM can be combined with any anomaly detection method to achieve better performance, the visualization results are presented in Fig. 3(d). By combining the anomaly detection and semantic segmentation results, the final segmentation prediction graph is shown in Fig. 3(e). In addition, we have tried to combine our method with some other uncertainty estimation methods and all have shown their effectiveness. Training a model to place attention on anomalous obstacles produces an interesting optimization problem where the network can gain rewards and reduce the overall loss if it succeeds in placing attention on its uncertainty region. Further details about training can be found in the following section: Experiments.

IV. Experiments

1) Implementation Details: We follow the experimental setup of SML [23]. In particular, DeepLabv3+ [38] with ResNet101 [39] backbone is selected as the segmentation module. The output stride is set to 8. We insert our attention module after the convolution layer following the concatenation operation of the decoder. The entire segmentation network is trained on the Cityscapes dataset [40], which is widely used in autonomous driving scenarios. By quantitative experiment, α in Eq. (3) initialized as 0 and the hyper-parameter τ and β in Eq. (4) and Eq. (7) are set to 50000 and 0.5. The anomaly score weighting hyper-meter σ is set to 20000.

2) Datasets: We evaluate the performance of our framework CuDAM on the standard benchmarks of anomaly or uncertainty estimation for urban driving: Fishyscapes (FS) Static, FSLost & Found [14] and Road Anomaly [12]. For all datasets, we provide evaluations on the public validation images. The FS Static and FS Lost& Found dataset separately contains 30 images with the unexpected obstacles from PASCAL VOC and 100 images regarding the obstacles in the original Lost & Found dataset. The Road Anomaly dataset has 60 images of a real scene from the Internet. Images in FS Static are obtained through the fusion of object and scene, while in FS Lost & Found, anomaly detection of small targets is more emphasized. Compared to the FS dataset, the unexpected objects in the Road Anomaly dataset contain a greater variety of species and size, and are therefore more challenging.

3) Evaluation Metrics: For the quantitative measure of our CuDAM approach's performance, we compute the average precision (AP), the false positive rate at 95% true positive rate (FPR95), and the area under receiver operating characteristics (AUROC) to validate our work. AP is an evaluation metric for measuring the accuracy of rare occurrences detection, which makes it suitable for anomaly detection. To highlight safety-critical applications, we also compute the FPR95, which describes the probability of an unexpected sample being misclassified into a normal sample when the OoD sample is correctly classified at 95%. Although some studies in [27], [41] have shown that AUROC is not suitable for anomaly detection because the anomaly data is similar to long-tailed distribution with class imbalance, we still calculate this metric as a reference.

4) Baseline: We compare our approach against the representative uncertainty estimation methods in the FS benchmark which do not require external dataset and training or designing additional network modules. We also compare our framework with approaches that add a new task but do not significantly increase the difficulty of training.

A. Evaluation Results

We evaluate our CuDAM technique by applying it to several different unexpected detection methods which do not require retraining or additional OoD data: MSP [21], Max logits [22] and SML [23]. CuDAM achieves a significant performance gain over using the SML. Tab. I shows comparisons between our proposed CuDAM approach and the baseline for the FS validation datasets and Road Anomaly. All data are publicly available instead of our own training results. For a fair comparison, we list the three criteria of the works: extra component, extra network, and utilizing OoD data. The extra component means that the method requires a modification to the original semantic segmentation model structure, which may lead to a worse semantic segmentation result. The extra network refers to the fact that the method requires the training of additional models (i.e., image synthesis, generative adversarial networks), which increases the inference time as well as the computational effort. Meanwhile, the use of additional datasets is not encouraged as it is difficult to collect enough data to cover all possible classes of anomalies in real scenarios for training purposes.

As shown in Tab. I, our method outperforms most other previous methods in FS Static and Road Anomaly datasets with a large margin. We still report methods that require additional networks or data. Methods that require reconstruction mostly require several times more computational cost and using additional OoD data is unrealistic. Our work aims to propose a component that can be plug-and-play on the original semantic segmentation network conversely. We notice that CuDAM could not obtain obvious promotion on the FS Lost & Found dataset compared to the other two datasets. Since obstacles in the FS Lost & Found dataset are usually placed far away, they became very small in the image after visualizing the failure cases. However, our approach tends to give high attention values to distant objects, which leads to interference with the result. This can be improved by adjusting the attention weights to different datasets, but this is not the emphasis of our study.

We then evaluate CuDAM on commonly used probability scores with the FS Lost & Found validation set. For requiring an external retraining methods, retraining may decrease the performance of the semantic segmentation. Therefore, we also report the mean intersection over union (mIoU). As can be seen from Tab. II, CuDAM performs particularly well without greatly increasing the computation and training complexity. This demonstrates the practicality of CuDAM since it both

This article has been accepted for publication in IEEE Transactions on Intelligent Vehicles. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIV.2022.3204714

TABLE I
RESULT COMPARISON ON THE FISHYSCAPES AND ROAD ANOMALY DATASETS

Methods	Additional Train		Utilizing OoD Data	FS Static			FS Lost & Found			Road Anomaly		
	Extra Component	Extra Network		AP ↑	FPR95↓	AUROC↑	AP ↑	FPR95↓	AUROC↑	AP ↑	FPR95↓	AUROC↑
MSP [21]	✘	✘	✘	14.24	34.10	88.94	6.02	45.62	86.99	20.59	68.44	73.76
Max logits [22]	✘	✘	✘	27.99	28.50	92.80	18.77	38.13	92.00	24.44	64.85	77.97
SML [23]	✘	✘	✘	48.67	16.75	96.69	36.55	14.53	96.88	25.82	49.74	81.96
Entropy [30]	✘	✘	✘	21.78	33.74	89.99	13.91	44.85	88.32	22.38	68.15	75.12
kNN Embedding-Density [14]	✘	✘	✘	4.10	22.30	-	-	-	-	-	-	-
Mahalanobis [42]	✘	✘	✘	27.37	11.7	96.76	56.57	11.24	96.75	14.37	81.09	62.85
Energy [43]	✔	✘	✘	31.66	37.32	91.28	25.79	32.26	93.50	24.44	63.36	78.13
Ours	✔	✘	✘	**66.26**	**5.93**	**98.61**	**36.20**	**29.50**	**95.02**	**30.28**	**48.71**	**84.29**
Synboost [27]	✘	✔	✔	48.44	47.71	92.03	40.99	34.47	94.89	41.83	59.72	85.25
SynthCP [15]	✘	✔	✔	23.22	34.02	89.90	6.54	45.95	88.34	24.86	64.69	76.08
Deep Gambler [44]	✘	✘	✔	67.69	15.39	97.51	39.77	12.41	97.19	31.45	48.79	85.45
Meta-OoD [27]	✔	✔	✘	72.91	13.57	97.56	56.57	11.24	96.75		-	-
Road inpainting [35]	✔	✔	✘	-	-	-	81.00	9.10	-	52.60	47.10	-
PEBAL [36]	✔	✘	✔	82.73	6.81	99.23	59.83	6.49	99.09	62.37	28.29	92.51

TABLE II
RESULTS WITH DIFFERENT UNCERTAINTY SCORES

Methods	mIoU	FS Static		
		AP ↑	FPR95↓	AUROC↑
MSP [21]	77.90	23.23	26.23	85.45
Max logits [22]	77.90	45.80	19.34	92.28
MSL [23]	77.90	52.30	11.25	96.59
CuDAM+MSP	78.29	36.67	19.16	91.25
CuDAM + Max logits	78.29	52.46	15.18	96.30
CuDAM+MSL	**78.29**	**66.26**	**5.93**	**98.61**

TABLE III
RESULTS OF ABLATION STUDY

Components of CuDAM			mIoU	FS Static		
Attention	Reward	L_{att}		AP ↑	FPR95↓	AUROC↑
✘	✘	✘	77.90	52.30	11.25	96.59
✔	✘	✘	+0.24	+2.33	-1.00	+1.05
✔	✔	✘	**+0.52**	+12.68	-3.97	+1.86
✔	✘	✔	+0.06	+3.93	-1.09	+0.54
✔	✔	✔	+0.39	**+13.96**	**-5.52**	**+2.02**

shows a positive impact on the semantic segmentation task and the anomaly detection task. Moreover, the models with CuDAM improve semantic segmentation by a small margin different from other approaches via introducing additional training tasks. Results demonstrate the general applicability of CuDAM across different uncertainty estimation methods. Researchers can seamlessly integrate CuDAM with anomaly detection approaches in any semantic segmentation network and jointly train the combined CuDAM-enhanced anomaly detection networks.

B. Ablation Study

We attempt to get a clearer understanding of how certain components of CuDAM contribute to the overall performance in this section. Tab. III describes the effect of each proposed method in CuDAM: attention module, reward mechanism, and L_{att}. In particular, since the calculations in the reward mechanism and L_{att} are based on attention module, so their effects on the overall performance cannot be discussed separately from the attention module.

We first find that using only the attention module, which computes an attention map of the same size as the input feature without applying any reward or loss function to constrain the generation of attention, can positively affect the performance. This result may be explained by the fact that the attention mechanism adopts the softmax function to control all values on the attention graph to sum to 1, which leads to competition among pixels to obtain a more accurate prediction of semantic segmentation, thus increasing the attention value of uncertain positions. For the three parts of CuDAM, the reward mechanism improves the overall performance most signification, showing a 10% increase in AP and a 3% decrease in FPR95. We also find that the model trained with L_{att} produces better accuracy in anomaly road obstacle detection while resulting in an unwanted decrease in mIoU. The decrease is common in multi-task learning with multiple loss function training and can be improved by adjusting the weighting parameter of the loss function.

Focus on a more intuitive perception of the role of each component, we also try to visualize the attention maps generated by models using different components. The visualization results are presented in Fig. 4. When the attention module is used alone, although it can achieve a certain effect of identifying unknown obstacles, the model tends to arrange the attention value as a small average value. The addition of the reward mechanism is similar to the addition of a regularization term which further enlarges the distance between the attention values of pixels with high uncertainty and low uncertainty. Compared with the attention-only model, the detection rate of unknown obstacles is significantly improved. It can be noticed that the addition of L_{att} makes the model hardly pay attention to its very definite region, and the effect of attention becomes more apparent. We observe that in some cases, the attention loss function will undesirably over-focus attention on small distance targets and ignore the nearby abnormal obstacles. We ultimately chose to retain the loss function because it helps to filter out most of the defined areas more easily by setting thresholds, which is worked for practical applications. Therefore, it can be demonstrated that our method significantly reduces the number of false pixels.

This article has been accepted for publication in IEEE Transactions on Intelligent Vehicles. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIV.2022.3204714

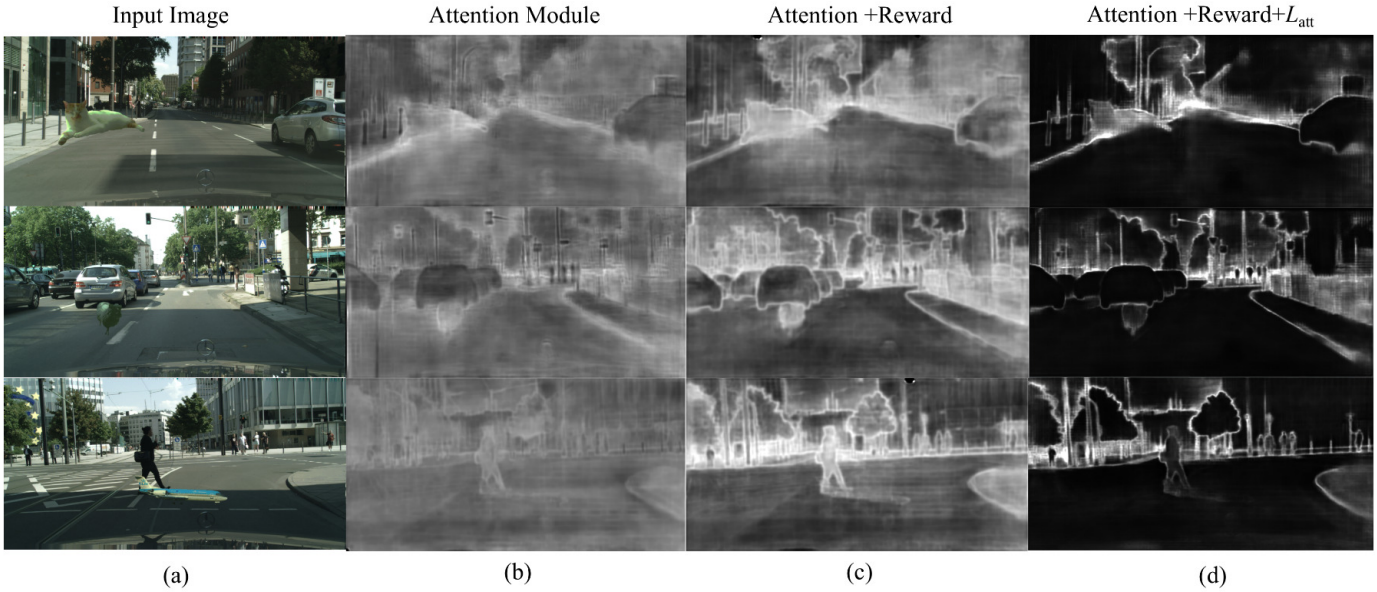

Fig. 4. The visualization results of **attention maps** generated by models using different components for ablation study. Each row corresponds to different input images, and the attention maps are shown as the three components (i.e., attention module, reward mechanism and L_{att}) are added to the model in turn. We show the output of attention maps directly without any threshold filtering that could embellish experiment results.

TABLE IV
COMPARISON ON COMPUTATIONAL COST

Models	GFLOPs	Infer. Time(ms)	Param.(M)
ResNet-101 [39]	2139.68	60.54	64.24
ResNet-101+SML [23]	2139.68	61.41	64.24
Ours	2140.19	67.23	64.58
SynthCP [15]	4551.11	146.90	-
Synboost [27]	-	1055.5	-

TABLE V
COMPARISON OF DIFFERENT ATTENTION MODULE POSITION

Attention Position	mIoU	FS Static		
		AP ↑	FPR95↓	AUROC↑
The surface position	78.43	63.17	6.93	97.71
The deep position	78.29	66.26	5.93	98.61

C. Comparison on computational cost

We report the computational overhead of the proposed CuDAM method in Tab. IV. CuDAM has a minimal computation cost of 0.51 GFLOPs and 5.82ms inference time, together with a 0.34M increase of parameter number. The model is implemented using PyTorch. Our metrics are measured with the image size of 2048 × 1024 and averaged over 100 runs on NVIDIA RTX3090. This result demonstrates that our method is lightweight and general, which can be integrated into most semantic segmentation modules. In particular, we also select SynthCP [15] and Synboost [27] as representative methods from the image re-synthesis methods and report their computational cost. The results show that our method significantly reduces computational loss compared to the synthetic method.

D. Arrangement of the Attention Module

After being given an attention module, inserting it into the different locations of the semantic segmentation network may affect the overall performance. In this experiment, we compare two different positions of arranging the attention module, i.e., the surface position near the predictive output layer and the deep position near the encoder. Empirically, the closer the attention is to the output, the greater the impact on the output results, which results in a more accurate attention map. On the other hand, the closer the attention is placed to the deeper position of the model for high-dimensional features, the more information is obtained and the greater the impact on the segmentation results. Tab. V shows a comparison of the experimental results for the two locations. From the result, we can find that placing the attention module in the deeper position infers a finer attention map than placing it near the output, i.e., in the training phase: insert location feature size is (304, 180, 180) and output feature size is (19, 720, 720). We consider that such experimental results may arise because the purpose of our curiosity attention is to focus on picture content rather than location. In layers close to the output, the attention weights extracted by the attention model are under-generalized, yet attention is sensitive, which tends to negatively affect the results. Conversely, when attention is placed in deeper layers, the accurate attention map causes the features in the uncertain regions of the model to be enhanced, and this enhancement is adequately fitted by the decoder, which can play the role of the attention module in enhancing local feature recognition. Meanwhile, this distinguishes CuDAM from uncertainty prediction tasks.

This article has been accepted for publication in IEEE Transactions on Intelligent Vehicles. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIV.2022.3204714

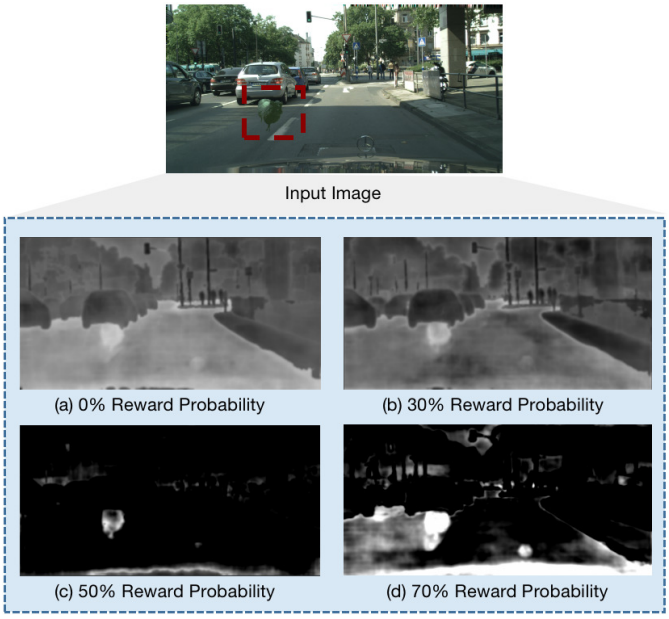

Fig. 5. The visualization results of **anomaly prediction outputs** with different reward probabilities. The 50% reward probability produces the lowest false positive rate of abnormal prediction output.

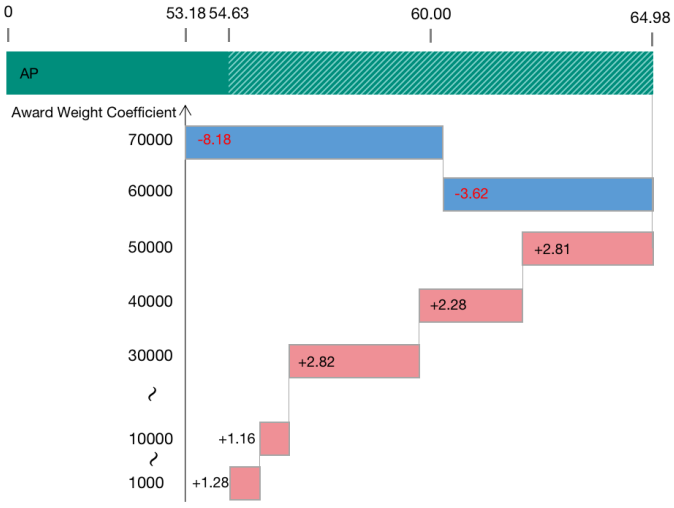

Fig. 6. Comparison of different reward weight parameters. We reflect both the AP change curve and the AP value in the same graph. The green bar represents the AP increasing from left to right, and the critical point corresponds to the initial value without the bonus mechanism. The pink and blue bar areas represent the magnitude of the change in AP value, with pink and blue corresponding to an increase in AP value and an undesired decrease, respectively.

E. Setting of the Reward Mechanism

Different rewards system settings will have different effects on anomaly detection results. The excessive reward can cause the model to learn how to model the decision lazily, which can affect the segmentation effect. While too limited rewards may lead to poor performance of anomaly detection. Therefore, in this section, we analyze the effect of different reward settings on the effectiveness of unknown anomaly detection. The reward mechanism we set is mainly reflected in two points: one is the reward probability, i.e., how much probability the model will be given a reward in the iteration; the other is the reward weight, i.e., how much correct information the model can get from the GT label.

We first search for an effective approach to set the possibility of giving a reward. We visualize the anomaly scores output by the model at different reward ratios, and Fig. 5 shows the visualization results. We observe that the model can already obtain better anomaly segmentation performance when the probability is set to 30%. However, increasing the reward proportion causes an increase in the uncertainty score at the edges of objects, which can interfere with the anomaly detection as in Fig. 5(b). When increasing the reward ratio to 50%, as shown in Fig. 5(c), the prediction output has the lowest false positive rate, and the model performs best in abnormal segmentation. Then continuing to increase the probability of giving a reward can cause the model to refuse to learn the decision boundary and lead to a problem similar to pattern collapse.

We also investigated the effect of the attention weights on the effect of the model. To intuitively display the segmentation performance and the magnitude of the effect on AP under different weights, we present the experimental results in the form of a statistical graph in Fig. 6. This result is significant at the reward weight $\tau = 50000$. Increasing the weight value further results in a very significant decrease in AP. Therefore, it can be assumed that our reward mechanism is effective in some cases, since focusing precisely on uncertain objects can maximize the reward obtained by the model. From the perspective of training, this can maximize the reduction of loss function for each forward propagation. However, too much weight means that the model is subjected to a strong regularization constraint, which can affect both the semantic segmentation and the anomaly detection undesirably. This produces an interesting optimization problem, where the model interacts between getting more rewards and getting more accurate semantic segmentation results. This case would drive the model to get more secure and reliable segmentation results.

V. Conclusion

In this paper, a CuDAM method has been proposed for pixel-wise unexpected road obstacle detection in complex driving scenes. We design an attention model stimulated by a reward mechanism and then design a loss function to further allocate attention weight. The attention graph generated by the above method is utilized to help the network determine anomalous obstacles. The final network learns where the model is unknown to emphasize and refines intermediate layer features effectively. Various experiments are conducted to validate the effectiveness of the method. Finally, we visualize how the network exactly infers under different reward settings. Although improving semantic segmentation performance is not the emphasis of this paper, it is interesting to note that experiments demonstrate such simple attention has a positive impact on semantic segmentation. In the future, we will directivity investigate the application of such a curiosity-driven attention to various additional tasks which is natural for the artificial neural network.

REFERENCES

[1] L. Chen, X. Hu, W. Tian, H. Wang, D. Cao, and F.-Y. Wang, "Parallel planning: A new motion planning framework for autonomous driving," *IEEE/CAA Journal of Automatica Sinica*, vol. 6, no. 1, pp. 236–246, 2018.

[2] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, 2019, pp. 3146–3154.

[3] Y. Kang, H. Yin, and C. Berger, "Test your self-driving algorithm: An overview of publicly available driving datasets and virtual testing environments," *IEEE Transactions on Intelligent Vehicles*, vol. 4, no. 2, pp. 171–185, 2019.

[4] S. Choi, J. T. Kim, and J. Choo, "Cars can't fly up in the sky: Improving urban-scene segmentation via height-driven attention networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9373–9383.

[5] Z. Zhang, Z. Mo, Y. Chen, and J. Huang, "Reinforcement learning behavioral control for nonlinear autonomous system," *IEEE/CAA Journal of Automatica Sinica*, vol. 9, pp. 1–13, 2022.

[6] G. Chen, K. Chen, L. Zhang, L. Zhang, and A. Knoll, "Vcanet: Vanishing-point-guided context-aware network for small road object detection," *Automotive Innovation*, vol. 4, no. 4, pp. 400–412, 2021.

[7] K. Gupta, S. A. Javed, V. Gandhi, and K. M. Krishna, "Mergenet: A deep net architecture for small obstacle discovery," in *2018 IEEE International Conference on Robotics and Automation*. IEEE, 2018, pp. 5856–5862.

[8] L. Sun, K. Yang, X. Hu, W. Hu, and K. Wang, "Real-time fusion network for rgb-d semantic segmentation incorporating unexpected obstacle detection for road-driving images," *IEEE Robotics and Automation Letters*, vol. 5, no. 4, pp. 5558–5565, 2020.

[9] Z. Hu, Y. Xing, W. Gu, D. Cao, and C. Lv, "Driver anomaly quantification for intelligent vehicles: A contrastive learning approach with representation clustering," *IEEE Transactions on Intelligent Vehicles*, 2022, doi:10.1109/TIV.2022.3163458.

[10] Y. Jin, X. Ren, F. Chen, and W. Zhang, "Robust monocular 3D lane detection with dual attention," in *Proceedings of 2021 IEEE International Conference on Image Processing*, 2021, pp. 3348–3352.

[11] C. Hubmann, J. Schulz, M. Becker, D. Althoff, and C. Stiller, "Automated driving in uncertain environments: Planning with interaction and uncertain maneuver prediction," *IEEE Transactions on Intelligent Vehicles*, vol. 3, no. 1, pp. 5–17, 2018.

[12] K. Lis, K. Nakka, P. Fua, and M. Salzmann, "Detecting the unexpected via image resynthesis," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 2152–2161.

[13] T. Vojir, T. Šipka, R. Aljundi, N. Chumerin, D. O. Reino, and J. Matas, "Road anomaly detection by partial image reconstruction with segmentation coupling," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 15 651–15 660.

[14] H. Blum, P.-E. Sarlin, J. Nieto, R. Siegwart, and C. Cadena, "The fishyscapes benchmark: Measuring blind spots in semantic segmentation," *International Journal of Computer Vision*, vol. 129, no. 11, pp. 3119–3135, 2021.

[15] Y. Xia, Y. Zhang, F. Liu, W. Shen, and A. L. Yuille, "Synthesize then compare: Detecting failures and anomalies for semantic segmentation," in *Proceedings of European Conference on Computer Vision*, 2020, pp. 145–161.

[16] K. Doshi and Y. Yilmaz, "Fast unsupervised anomaly detection in traffic videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 624–625.

[17] K. Kobayashi, S. Ravaioli, A. Baranès, M. Woodford, and J. Gottlieb, "Diverse motives for human curiosity," *Nature Human Behaviour*, vol. 3, no. 6, pp. 587–595, 2019.

[18] A. Kendall, V. Badrinarayanan, and R. Cipolla, "Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding," *arXiv preprint arXiv:1511.02680*, 2015.

[19] B. He, B. Lakshminarayanan, and Y. W. Teh, "Bayesian deep ensembles via the neural tangent kernel," *Advances in Neural Information Processing Systems*, vol. 33, pp. 1010–1022, 2020.

[20] A. Kendall and Y. Gal, "What uncertainties do we need in Bayesian deep learning for computer vision?" *Advances in Neural Information Processing Systems*, vol. 30, pp. 5579–5584, 2017.

[21] D. Hendrycks and K. Gimpel, "A baseline for detecting misclassified and out-of-distribution examples in neural networks," *arXiv preprint arXiv:1610.02136*, 2016.

[22] D. Hendrycks, S. Basart, M. Mazeika, M. Mostajabi, J. Steinhardt, and D. Song, "Scaling out-of-distribution detection for real-world settings," *arXiv preprint arXiv:1911.11132*, 2019.

[23] S. Jung, J. Lee, D. Gwak, S. Choi, and J. Choo, "Standardized max logits: A simple yet effective approach for identifying unexpected road obstacles in urban-scene segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 15 425–15 434.

[24] C. Wang, W. Pedrycz, Z. Li, and M. Zhou, "Residual-driven fuzzy C-means clustering for image segmentation," *IEEE/CAA Journal of Automatica Sinica*, vol. 8, no. 4, pp. 876–889, 2020.

[25] J.-A. Bolte, M. Kamp, A. Breuer, S. Homoceanu, P. Schlicht, F. Huger, D. Lipinski, and T. Fingscheidt, "Unsupervised domain adaptation to improve image segmentation quality both in the source and target domain," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 1404–1413.

[26] C. Creusot and A. Munawar, "Real-time small obstacle detection on highways using compressive rbm road reconstruction," in *Proceedings of 2015 IEEE Intelligent Vehicles Symposium*, 2015, pp. 162–167.

[27] G. Di Biase, H. Blum, R. Siegwart, and C. Cadena, "Pixel-wise anomaly detection in complex driving scenes," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 16 918–16 927.

[28] T. Wang, X. Xu, F. Shen, and Y. Yang, "A cognitive memory-augmented network for visual anomaly detection," *IEEE/CAA Journal of Automatica Sinica*, vol. 8, no. 7, pp. 1296–1307, 2021.

[29] D. Hendrycks, M. Mazeika, and T. Dietterich, "Deep anomaly detection with outlier exposure," *arXiv preprint arXiv:1812.04606*, 2018.

[30] R. Chan, M. Rottmann, and H. Gottschalk, "Entropy maximization and meta classification for out-of-distribution detection in semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 5128–5137.

[31] P. Bevandić, I. Krešo, M. Oršić, and S. Šegvić, "Simultaneous semantic segmentation and outlier detection in presence of domain shift," in *Proceedings of German Conference on Pattern Recognition*, 2019, pp. 33–47.

[32] D. Hendrycks, M. Mazeika, and T. Dietterich, "Deep anomaly detection with outlier exposure," *arXiv preprint arXiv:1812.04606*, 2018.

[33] Y. Tian, X. Li, K. Wang, and F.-Y. Wang, "Training and testing object detectors with virtual images," *IEEE/CAA Journal of Automatica Sinica*, vol. 5, no. 2, pp. 539–546, 2018.

[34] T. Ohgushi, K. Horiguchi, and M. Yamanaka, "Road obstacle detection method based on an autoencoder with semantic segmentation," in *Proceedings of the Asian Conference on Computer Vision*, 2020, pp. 223–238.

[35] K. Lis, S. Honari, P. Fua, and M. Salzmann, "Detecting road obstacles by erasing them," *arXiv preprint arXiv:2012.13633*, 2020.

[36] Y. Tian, Y. Liu, G. Pang, F. Liu, Y. Chen, and G. Carneiro, "Pixel-wise energy-biased abstention learning for anomaly segmentation on complex urban driving scenes," *arXiv preprint arXiv:2111.12264*, 2021.

[37] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 3–19.

[38] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 801–818.

[39] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[40] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3213–3223.

[41] N. V. Chawla, N. Japkowicz, and A. Kotcz, "Special issue on learning from imbalanced data sets," *ACM SIGKDD Explorations Newsletter*, vol. 6, no. 1, pp. 1–6, 2004.

[42] K. Lee, K. Lee, H. Lee, and J. Shin, "A simple unified framework for detecting out-of-distribution samples and adversarial attacks," *Advances in neural information processing systems*, vol. 31, pp. 7167–7177.

[43] W. Liu, X. Wang, J. Owens, and Y. Li, "Energy-based out-of-distribution detection," *Advances in Neural Information Processing Systems*, vol. 33, pp. 21 464–21 475, 2020.

[44] Z. Liu, Z. Wang, P. P. Liang, R. R. Salakhutdinov, L.-P. Morency, and M. Ueda, "Deep gamblers: Learning to abstain with portfolio theory," *Advances in Neural Information Processing Systems*, vol. 32, pp. 10 623–10 633, 2019.

This article has been accepted for publication in IEEE Transactions on Intelligent Vehicles. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIV.2022.3204714

Xiangxuan Ren received the B.S. degree in automation from University of Electronic Science and Technology of China, Chengdu, China, in 2020. She is currently pursuing the M.S. degree in electronic information at the Department of Automation, Shanghai Jiao Tong University, Shanghai, China.

Her current research interests include the perception of autonomous vehicles, pattern recognition, weakly supervised learning, anomaly detection and sensor fusion.

Min Li received the B.S. degree in electrical engineering and automation from Shanghai University of Electric Power, China, in 2020. She is currently pursuing the M.S. degree in electronic information at the Department of Automation, Shanghai Jiao Tong University, Shanghai, China.

Her research interests include object detection, semantic segmentation and sensor fusion.

Zhenhua Li received the B.S. degree in automation from Harbin University of Science and Technology, Harbin, China, in 2017, and the M.S. degree in control theory and control engineering from Northeastern University, Shenyang, China, in 2020. He is currently pursuing the Eng.D. degree in electronic information at the Department of Automation, Shanghai Jiao Tong University, Shanghai, China.

His research interests include nonlinear control, observer theory, switched systems and time-delay systems.

Wentao Wu (Student Member, IEEE) received the B.E. degree in electrical engineering and automation from Harbin University of Science and Technology, Harbin, China, in 2018. He received the M.E. degree in electrical engineering from Dalian Maritime University, Dalian, China, in 2021. He is pursing the Ph.D. degree in electronic information from Shanghai Jiao Tong University, Shanghai, China.

His current research interests include unmanned surface vehicles, formation control, safety control, prescribed performance control.

Lin Bai received the M.S. degree in Circuit and System from Yanshan University, Hebei, China, in 2004. He is currently pursuing the Ph.D. degree with the Computer Department of Shanghai Jiao Tong University, Shanghai, China.

His main research interests include low-speed unmanned driving and cloud-edge-terminal integrated security patrol robots. He is the founder of a robot company and an expert member of TC591 of the China Robotics Standardization Committee.

Weidong Zhang (Senior Member, IEEE) received the B.S., M.S., and Ph.D. degrees from Zhejiang University, Hangzhou, China, in 1990, 1993, and 1996, respectively, and then worked as a Postdoctoral Fellow at Shanghai Jiaotong University.

He joined Shanghai Jiao Tong University, Shanghai, China, in 1998 as an Associate Professor and has been a Full Professor since 1999. He worked with the University of Stuttgart, Stuttgart, Germany, as an Alexander von Humboldt Fellow from 2003 to 2004. He is a recipient of National Science Fund for Distinguished Young Scholars, China and Cheung Kong Scholar, Ministry of Education, China. He is currently Director of the Engineering Research Center of Marine Automation, Shanghai Municipal Education Commission, and Director of Marine Intelligent System Engineering Research Center, Ministry of Education, China. His research interests include control theory, machine learning theory, and their applications in industry and autonomous systems. He is the author of 265 SCI papers and 1 book, and holds 72 patents.