Received 10 January 2024, accepted 23 January 2024, date of publication 21 February 2024, date of current version 27 February 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3359695

RESEARCH ARTICLE

A Novel Method for Road Anomaly Objects Detection in the Traffic Environment With Multi-Mechanism Fusion

WENYAN CI[1], JIAYIN XUAN[1], RUNZE LIN[2], AND SHAN LU[3]
[1]School of Engineering, Huzhou University, Huzhou 313000, China
[2]State Key Laboratory of Industrial Control Technology, Zhejiang University, Hangzhou 310027, China
[3]Institute of Intelligence Science and Engineering, Shenzhen Polytechnic, Shenzhen 518055, China

Corresponding author: Shan Lu (lushan@szpt.edu.cn)

This work was supported in part by the Huzhou City Public Welfare Technology Application Research Project under Grant 2022GZ10; in part by the Zhejiang Provincial Public Welfare Technology Application Research Project, China, under Grant LGG22F030016; and in part by the Postgraduate Research and Innovation Project of Huzhou University under Grant 2023KYCX57.

ABSTRACT In the modern automotive industry, Advanced Driving Assistance Systems (ADAS) have gradually become a standard feature in various types of vehicles, with the important function of detecting road anomalies. The appearance of anomalies on the road can be attributed to unexpected situations while driving, and the current methods for detecting distant or small anomalies are not highly accurate. Therefore, in this paper, a method is proposed that uses semantic segmentation to extract key features from the image, and obtaining a new synthesized image by image resynthesis. Then, segmentation uncertainty and depth information are used to compare the differences between multiple feature maps and the input image to highlight the anomalies. Additionally, a postprocessor is designed to use an anomaly score to enhance the recognition of anomaly target and reduce false positives caused by noise. Experiments are conducted on the Obstacle Track dataset and the Lost and Found dataset, and various methods for detecting anomaly objects are compared. The experimental results demonstrate that the method proposed in this paper can effectively detect un-common objects in the training dataset in road anomaly object detection. It improves the detection rate and reduces the false positive rate based on previous anomaly detection methods. The proposed method presented in this paper achieves high detection rates for both seen and unseen anomaly objects in the training set, which enhances the generalization ability of anomaly detection in the road area of interest.

INDEX TERMS ADAS, depth information, image resynthesis, postprocessor, semantic segmentation.

I. INTRODUCTION

With the increasing number of cars on the road, the need for safe driving has become crucial [1]. Advanced Driving Assistance System (ADAS) has emerged as a mainstream direction in the development of automotive safety [2]. Its main task is to detect road anomaly objects and provide timely feedback to prevent accidents. The anomaly objects here refer to objects that are uncommon and complex in shape [3], [4] in the road environment, so detecting anomaly objects on the road is a difficult task. Deep learning is currently a commonly used method for complex object detection tasks. However, existing networks are mainly trained to identify targets by known categories, and it is difficult to detect anomaly objects outside the existing categories in the dataset. This will cause the generalization of the perception system to deteriorate, making the system unable to accurately identify anomaly targets appearing on the road, which is a very dangerous situation. Therefore, finding an effective method to improve the detection rate of anomaly targets in ADAS is a major challenge.

Traditional semantic segmentation techniques, exemplified by state-of-the-art PSPNet [5] and DeepLabv3+ [6], exhibit remarkable performance when applied to image datasets with known classes, producing highly accurate segmentation results. However, when encountering certain

The associate editor coordinating the review of this manuscript and approving it for publication was Shovan Barma.

anomaly objects such as cargo dropped from a vehicle in front or rocks unexpectedly standing in the middle of the road, these models may confuse anomaly objects with the environment and failing to recognize them as anomaly. This can have fatal consequences in autonomous driving scenarios, where failure to identify anomaly objects may lead to collisions with such objects. In light of this, this paper explores two existing methods to develop a novel anomaly detection model. The first method involves lever-aging uncertainty estimation [7], [8] and depth input [9], [10] to detect anomalies, collectively referred to as assistant methods. The uncertainty estimation method differs from semantic segmentation in its inclination to predict low-confidence scenarios as anomalies, thereby estimating higher uncertainty of unknown classes in images. However, this method is prone to interference from unknown noise, resulting in a relatively high false positive rate. The depth input method extracts features from the depth map and performs feature fusion with the original image, highlighting geometric elements in the scene and providing richer semantic information. However, this method still lacks the ability to accurately distinguish between known classes and anomalies. The second method involves resynthesizing a new input image from the semantic map predicted by the segmentation network, and detecting anomalies by analyz-ing the feature differences among the original input image, the predicted semantic map, and the synthesized image [11]. This method exhibits a high extraction effect when dealing with objects outside the network's training classification but introduces a problem that cannot be ignored. This model is overconfident in predicting features beyond the network's training classification, which results in a consid-erable impact of unknown noise on the model's prediction results. The model treats unknown noise as an anomaly target for feature extraction, which does not represent the anomaly difference between the input image and the syn-thetic image desired and renders the comparison more complicated.

In this paper, a novel anomaly object detection method is proposed, which combines the assistant method and the resynthesis method to achieve more accurate detection of anomaly objects and reduce the impact of unknown noise on the difference comparison. The proposed method is demonstrated through experiments, and it is found that the fusion of the assistant method and the resynthesis method can effectively improve the detection effect. Moreover, the two methods complement each other to solve as many anomaly scenarios as possible. Furthermore, a postprocessor is proposed in this paper to enhance the localization of road anomalies. This module uses a linear clustering meth-od to segment road images into superpixels and calculates the anomaly score of each superpixel to more accurately locate the anomalies on the road. By designing the calcula-tion formula of the anomaly score, the module can better distinguish anomalies on the road from other irrelevant objects, thereby reducing the occurrence of false positives. The main contributions of this paper are as follows:

1) A novel anomaly detection method is proposed. Traditional segmentation methods are unable to detect objects outside the distribution of the training set, and thus are inadequate for anomaly detection. To overcome this limitation, our approach leverages the complementary nature of assistant and resynthesis methods to mitigate the influence of unknown noise during the comparison of dis-similarity networks. This targeted approach to difference detection results in more effective detection of anomaly objects.

2) A postprocessor is proposed. This method performs superpixel segmentation on the anomaly detection map output by the anomaly detection model, and outputs an anomaly score map to strengthen the positioning of anomaly objects in the input image. By designing the anomaly scoring formula, it pays more attention to the anomalies on the road, and tries to ignore the anomalies in other scenes to reduce false positives.

II. RELATED WORK

During the early stages of the development of autonomous driving technology, sample division of known datasets was used for feature extraction and classification. Shallow classifiers were then utilized to learn artificial features, enabling the detection and identification of anomaly objects on the road. However, the limitations of detection-intensive and classification capabilities of such methods of shallow classifiers have rendered them unable to meet the current requirements for road anomaly objects detection tasks. In recent years, detection methods based on deep learning have witnessed a significant advancement. Convolution Neural Network (CNN) has been found to have a strong ability to extract features and has been able to complete the required classification tasks by learning a large amount of data. Consequently, deep learning has emerged as a powerful tool, which has been applied to the field of road anomaly objects detection with remarkable success.

A. SEMANTIC SEGMENTATION

Semantic segmentation algorithms are often used in the field of image detection and classification, which can make dense predictions for each pixel, thereby achieving pixel-by-pixel category labeling. As the earliest representative of semantic segmentation, FCN [12] (Full Convolutional Network) has the advantage that it is not affected by the size of the input image, and replaces the fully connected layer in the network structure with a skip layer and a deconvolution layer. In this way, the pixel-level segmentation of the image is realized. SegNet [13] was proposed in 2015 based on FCN. SegNet utilizes the first 13 layers of the VGG16 convolutional network as an encoder, each encoder layer corresponds to the decoder layer one by one, and the decoder's outputs are processed by the softmax classifier, and finally independently generates class probabilities for each pixel. Paszke et al. [14] proposed Enet, a real-time deep neural network that preserves segmentation accuracy, reduces parameter quantity, and improves operating speed, making it applicable to embedded devices. In 2018, Chen et al. [6] proposed the DeeplabV3+

segmentation model, which combines the encoding-decoding structure with ASPP, and introduces dilated convolution to expand the receptive field of the model, thereby enhancing the ability of the model to segment targets of different sizes. Valdez-Rodríguez et al. [15] combined the advantages of semantic segmentation to obtain local information with the depth estimation method, and used a mixed dataset for training. A 2D-3D hybrid CNN network is proposed, which can estimate the depth of a single image and segment the objects found in it. A method proposed by Wang et al. [16] integrates different visual features with semantic segmentation. In their work, a transformed disparity image is introduced, which makes the values of the drivable region similar, while highlighting the significant differences between the drivable region and road anomalies/damage. This aids in distinguishing between the drivable area and road anomalies. However, in order for the semantic segmentation model to recognize all categories during testing, the model must have encountered these categories during training. This is obviously unrealistic in some complex scenes, such as anomaly objects that suddenly appear in road traffic scenes. These anomaly objects, since they have not appeared during model training, are likely to be predicted as known categories or not predicted at all, which would be fatal for autonomous driving scenarios.

B. UNCERTAINTY ESTIMATION

With the increasing focus on anomaly object detection in scenes, there has been a growing interest in reasoning about uncertainty in neural networks among scholars. Uncertainty in neural networks can be measured using probabilities from a softmax distribution, and samples can be classified as out-of-distribution using simple statistics, as proposed by Hendrycks et al. [17]. In practical applications, dropout [18] is a common approximate Bayesian inference method [19], which has been widely used in the field of semantic segmentation, such as in Bayesian SegNet [20] and its extension work [8]. Isobe [21] combined uncertainty thresholds with Bayesian SegNet to distinguish erroneous regions in a scene. However, this approach often yields high uncertainty estimates at object boundaries, as it cannot definitively assign any label to the object, thereby failing to predict the expected anomaly objects. The problem of target object boundary in anomaly objects detection was solved by Mukhoti et al. [22], who used a Bayesian neural network with MC dropout to estimate the uncertainty of pixels. They distinguished between accidental and cognitive uncertainty, but pixel-level detection of anomaly target objects is not accurate as a whole. Rottmann et al. [23] predicted some possible regions of high error by aggregating different discrete measures, such as the differ-ence in entropy and softmax probability. This alleviated the problem of predicting object boundaries. Oberdiek et al. [24] demonstrated that visual feature differences can be exploited to identify anomalies in high-error regions by detecting and retrieving objects that are not within the distribution of the training set in semantic segmentation. Vojíř et al. [31] achieved good performance using a unique autoencoder-like architecture, image conditional distance features, and drawing modules. Gudovskiy et al. [32] adopt a standard flow framework to improve the robustness of semantic segmentation models in real data environments with distribution shifts and outliers. Simultaneous intradistribution misclassification (IDM) and outlier class detection are then implemented via energy input to achieve a low-complexity 2D architecture without the need for tedious retraining of pre-trained semantic segmentation models. Despite these advancements, accurately localizing anomalies remains a challenge, resulting in many false positive predictions. As a result, such methods are often prone to failure in road anomaly detection tasks.

C. DISSIMILARITYT DETECTION

At present, there is a new idea to detect anomaly objects in the scene. The input image is resynthesized by synthesis model. There will be an anomaly appearance difference between the resynthesized image and the input image, so the anomaly appearance difference can be used to locate the anomaly objects. In early work, autoencoders were generally used to resynthesize images [25], but the quality of resynthesized images by this method was poor. A reconstruction module was proposed by Vojir et al. [26] to identify and reconstruct road surfaces. In their work, the reconstruction module generates reconstruction errors and is coupled with semantic segmentation using trainable coupling blocks. This integration incorporates information from known classes and generates the final per-pixel anomaly scores for anomaly identification. With the rise of generative adversarial networks, new methods [27], [28], [29], [30] have utilized these networks to create new input images based on the semantic feature maps generated by the semantic segmentation model, making better use of the feature differences between input images and synthesized images. The advantage of dissimilarity network methods lies in not excessively relying on the segmentation quality of the segmentation network. The work in this paper proves that feeding uncertainty information and depth information as attention to the dissimilarity network can improve the detection effect of the model on anomaly objects in the scene.

III. METHODOLOGY

This paper presents a deep learning framework for detecting road anomaly objects. Initially, the image is processed for semantic segmentation, analyzed and calculated through a segmentation network, resulting in three output maps, including a semantic map and two uncertainty maps. The predicted semantic map is fed into the synthesis network, which generates a map that is highly similar to it. By calculating the feature differences between the generated image and the input image, the perceptual difference can be determined. To enhance the detection capabilities of the model, a new depth image is introduced and the RGB-D [9] network is employed to extract the features of the input image and depth

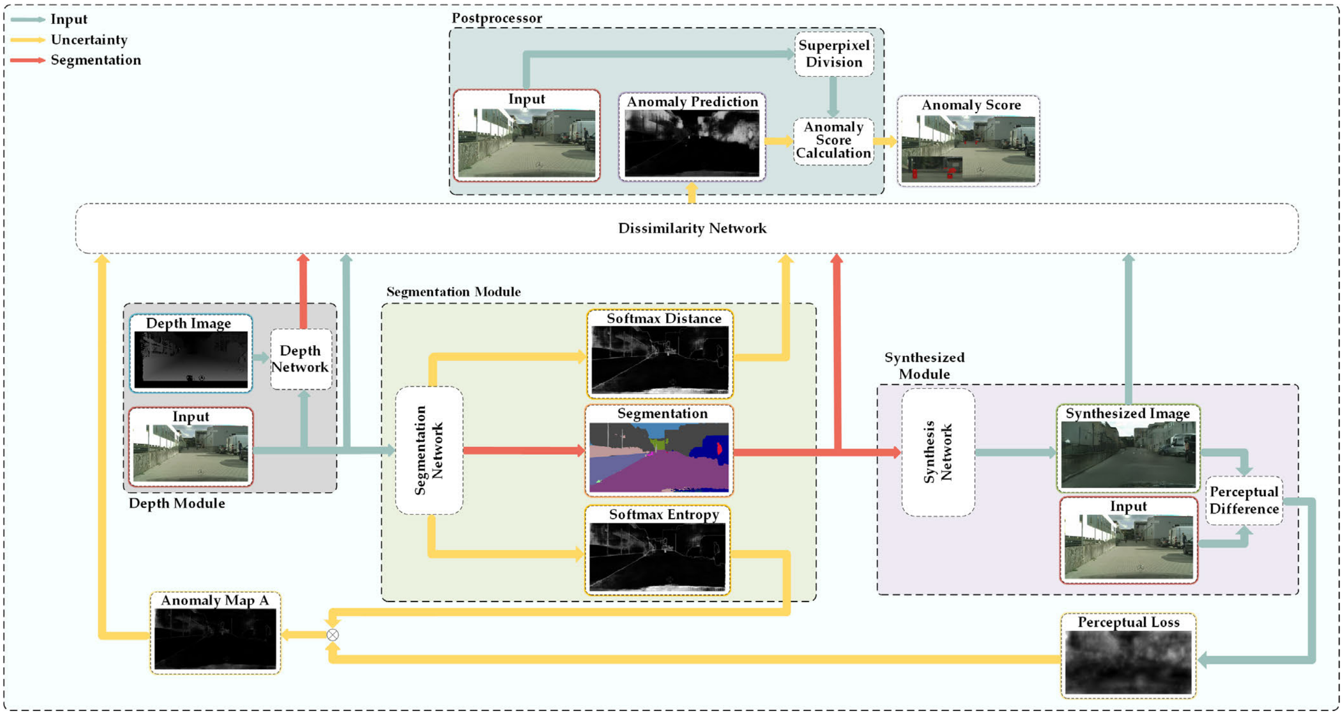

FIGURE 1. Flow chart of the method.

image separately. The channel attention mechanism is then applied to process the feature maps of the two branches, which are fused into new feature maps. Subsequently, the dissimilarity network trains on all predicted and input images, the anomaly prediction map is obtained. Finally, the anomaly map of each local region is sharpened by a postprocessor, and an anomaly score map is output. The flowchart of our proposed method is shown in Figure 1.

A. ANOMALY DETECTION MODULE

The road anomaly object detection module operates by segmenting the input image, the resulting semantic map is sent to a synthesis network for resynthesis, and using a dissimilarity network to compare the input image with the synthesized image to detect anomalies. On this basis, this paper incorporates new depth maps and improved uncertainty maps as assistant maps to enhance the module's ability to detect road anomalies. The module consists of four submodules, namely, the segmentation module, synthesis module, depth module, and dissimilarity module.

1) SEGMENTATION MODULE

The input image is fed into the segmentation network [6] to acquire the semantic map in the segmentation module. In addition to the semantic map, two dispersion measurements are also calculated to quantify the uncertainty in the semantic map predictions. The two dispersion measurements are the softmax entropy U and the softmax distance D (i.e., the difference between the two largest softmax values), which have been demonstrated to be effective in understanding errors within segmentation in the literature [23]. To calculate these measurements for each pixel x, the following equations are utilized:

$$U_x = -\sum_{c \in classes} p(c) log_2 p(c) \tag{1}$$

$$D_x = 1 - \max_{c \in classes} p(c) + \max_{c' \in classes \backslash (\arg\max_c p(c))} p(c') \tag{2}$$

Among them, $p(c)$ is the softmax probability of class c. Normalize both quantities to [0, 1].

2) SYNTHESIS MODULE

The synthesis module has the capability to generate a realistic image by taking a semantic mapping as input, where there exists pixels to pixels correspondence. In order to achieve this, the module is trained with a conditional generative adversarial network (cGAN) [33], [34] so that the semantic distribution of the input images can be effectively matched to the distribution of the generated images. The GAN network has achieved impressive progress in generating realistic urban scenes. However, the semantic map generated by the network misses essential color and appearance information, resulting in the inability to perfectly restore the image in terms of these features. Thus it can be compared pixel by pixel. To address this limitation, a perceptual loss method was proposed in [29] that compares objects by computing the perceptual difference between the original image and the synthesized image, rather than relying on low-level features such as color and texture. This method utilizes ImageNet pretrained VGG as a feature extractor, extracts features from it and finds the pixel with the largest feature difference. By detecting anomaly objects or misclassifications, erroneous feature representations of synthesized images can be identified. Perceptual differences should therefore be more sensitive to these differences. For each pixel x of the input image and the corresponding pixel r of the synthesized image, the perceptual loss is computed as follows:

$$L(x, r) = \sum_{i=1}^{4} \left\| F^{(i)}(x) - F^{(i)}(r) \right\|_1 \tag{3}$$

Among them, $F^{(i)}$ is the output of the i-th feature layer of the VGG network. Kamoi et al. [35] found that the performance of the dissimilarity network to detect anomaly objects is related to the selection of feature layers, and deeper feature layers may lose anomaly objects. Therefore, this paper selects the output of the first 4 layers of feature layers, and normalizes this dispersion measurements to [0, 1].

The semantic mapping of anomaly objects often contains ambiguity, which results in significant differences between the resynthesized image from the semantic mapping and the input image. To quantify this ambiguity, we use softmax entropy U as a measurement, while perceptual loss L is used to measure appearance difference. By multiplying these two measurements, an anomaly map A with deep features can be generated.

$$A = U \otimes L \tag{4}$$

3) DEPTH MODULE

The depth module leverages both the appearance information of the RGB image and the position and contour information of the depth map as its input. The feature extraction is performed using ResNet-18 [36] as the backbone network, as shown in Figure 2. Two branches are utilized to extract the input image's features, one for the RGB image branch and the other for the depth image branch. Attention Feature Complementary (AFC) module [9] is applied to fuse the features of these two branches at each layer of ResNet-18. Finally, the spatial feature pyramid pooling (SPP) module is utilized to generate multiscale feature maps with detailed information, which are then restored to their original resolution via upsampling.

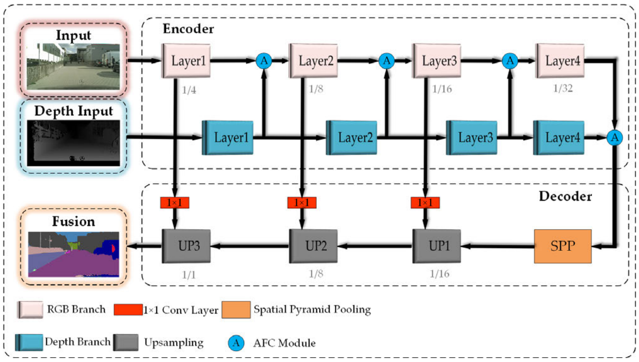

FIGURE 2. RGB-D network structure diagram.

The mentioned structure yields two feature maps, namely the RGB branch feature map $O_{in} = [O_{in1}, \cdots, O_{inC}] \in \mathbb{R}^{C \times H \times W}$ and the depth input branch feature map $D_{in} = [D_{in1}, \cdots, D_{inC}] \in \mathbb{R}^{C \times H \times W}$. Two channel attention mechanism modules are introduced in the AFC module to process the feature maps of the two branches respectively. It processes the feature map as a channel descriptor using global average pooling, adds a 1×1 convolutional layer with the same channel for full connection, and the convolution result is activated through the sigmoid function. The value of the weight matrix is then limited between 0 and 1. After obtaining the attention weight matrix of the two branches, the outer product is performed with the corresponding input feature map. The results of the two branches are added to obtain the final fusion feature map $F \in \mathbb{R}^{C \times H \times W}$, expressed as follows:

$$F = O_{in} \otimes \sigma_1[\phi_1(O_{in})] + D_{in} \otimes \sigma_2[\phi_2(D_{in})] \tag{5}$$

Among them, $\otimes$ represents the outer product, ϕ represents the calculation of global average pooling and 1×1 convolution, and σ represents the sigmoid function. By introducing the channel attention mechanism in both the RGB input branch and the depth input branch, the feature maps containing more informative features can obtain higher weight values. This allows the basic segmentation method to take advantage of the additional information provided by the depth map, thereby improving the accuracy of segmentation. The integration of depth information is notebly helpful in dissimilarity networks.

4) DISSIMILARITY MODULE

This module employs a variety of input features to predict anomaly segmentation maps. These input features encompass original images, synthesized image, semantic images with depth input, semantic images, as well as uncertainty maps (e.g. softmax distances and anomaly maps) computed by the segmentation network. By utilizing a dissimilarity network, these features can be effectively integrated and leveraged to enhance the predictive performance of anomaly segmentation maps. The network structure is shown in Figure 3.

The dissimilarity module comprises two encoder modules, one fusion module, and three decoder modules. The first decoder shares the same structure as VGG16, with three max-pooling layers that output a feature map after each layer. These feature maps are combined with the final feature map outputted by the encoder, resulting in a total of four feature maps. The encoder shares the weight used for encoding the original input image and the synthesized image. On the other hand, the second encoder consists of a 7×7 convolutional layer and three 3×3 convolutional layers, a feature map is output after each convolutional layer. This encoder is employed to encode original semantic information, deep input semantic information, and uncertainty information, with different weights used to encode these three types of information. The fusion module connects the feature maps of the input image,

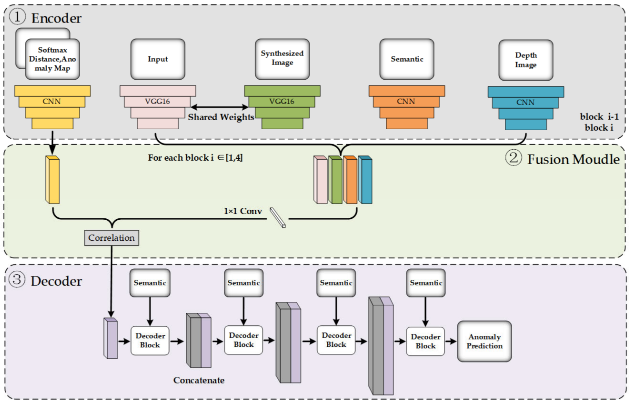

FIGURE 3. Dissimilarity module network structure.

synthesized image, semantic image, and deep input semantic image at different resolutions. It employs a 1 × 1 convolutional layer to fully connect the feature maps and outputs a feature map at each resolution. Finally, the result map of the 1 × 1 convolution is correlated point-by-point with the above discrete feature map. Four decoders are used in the dissimilarity network. The first and second decoders have the same structure, consisting of two 3 × 3 convolutional layers with the same number of filters, two SPADE normalized SELU layers, and a 2 × 2 transposed convolutional layer. The third encoder has two 3 × 3 convolutional layers with different numbers of filters, two SPADE normalized SELU layers, and a 2 × 2 transposed convolutional layer. The fourth encoder comprises two 3 × 3 convolutional layers with different numbers of filters, two SPADE normalized SELU layers, and a 1 × 1 transposed convolutional layer. The lowest resolution feature map is obtained by the first encoder, while the feature map of the fusion module is obtained by the second encoder. The output is then formed by concatenating the results from the first encoder with the feature map from the fusion module. This process continues for subsequent encoders.

B. POSTPROCESSOR

Through the utilization of dissimilarity networks for prediction, the model is now capable of accomplishing the localization of anomaly objects to a significant extent. However, due to limitations in the performance of the segmentation and synthesis networks, the dissimilarity network is highly susceptible to unnecessary noise interference when dealing with the feature disparities between the original and synthesized images. Consequently, these noises are erroneously identified as anomaly objects by the anomaly detection module, thereby elevating the false positive rate of the model's predictions. Consequently, the introduction of a Postprocessor is deemed necessary. The refinement of anomaly object localization is accomplished by the postprocessor. The anomaly prediction map is subjected to segmentation into superpixels, effectively capturing the spatial structure of anomalies. Subsequently, anomaly scores are assigned to each superpixel, leveraging their unique characteristics and anomaly prediction information. The outcome is an anomaly score map. The postprocessor is composed of a superpixel segmentation module and an anomaly score calculation module.

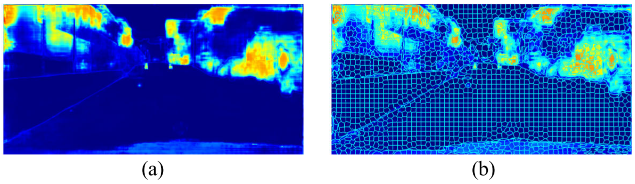

FIGURE 4. Superpixel segmentation results: (a) Anomaly prediction map; (b) Superpixel segmentation image.

1) SUPERPIXEL SEGMENTATION

Superpixel segmentation technique can classify pixels with adjacent positions and similar features such as color, texture, and brightness into small regions, which can be used to process images. It can also highlight object boundaries and improve the segmentation accuracy of smaller pixel regions in images. The implementation involves transforming the image into the CIELAB color space and combining the color value and position information of each pixel into a 5-dimensional vector. Based on the clustering approach, a set of seed points can be generated, and the pixels closest to each seed point can be grouped together until all pixels are classified. Then, the average vector value of each superpixel is calculated to obtain a new set of clustering centers, and the neighboring pixels are searched again until the final convergence. The anomaly prediction map output by the dissimilarity network is shown in Figure 4(a), while the corresponding superpixel segmentation image is shown in Figure 4(b). As shown in Figure 4(b), the scene is divided into many grids after superpixel segmentation, and the object boundaries in the scene are clearly distinguished, reducing the processing range and using more refined features for subsequent processing. The superpixel segmentation process is shown in Algorithm 1.

2) ANOMALY SCORE CALCULATION

In this paper, a novel anomaly score calculation method was proposed. This method takes the anomaly prediction output of dissimilarity network as input. Based on this input, we define the anomaly object's score in the i-th superpixel as follows:

$$S_i = \alpha_i p_j \sum_j n_j \exp(-\frac{r_{i,j}^2}{2\omega^2}) \tag{6}$$

Among them, α_i is the average of the anomaly scores in the j-th superpixel, $r_{i,j}$ is the Euclidean distance between the center position of the i-th superpixel and the center position of the j-th superpixel, ω is the median of the Euclidean distances between the center positions of each pair of superpixels. The final score is normalized to [0, 1]. Then, by setting a threshold, the calculated value S_i is compared to the threshold, and the region exceeding the threshold is recognized as the

Algorithm 1 Superpixel Segmentation

Input: k: The number of desired clusters; S: The reglar grid step size

Output: I: A segmented image

```
/* Initialization */
Initialize cluster centers $C_k = [r_k, g_k, b_k, x_k, y_k]^T$ by sampling pixels at regular grid steps $S$.
Move cluster centers to the lowest gradient position in a $3 \times 3$ neighborhood.
Set label $r(i) = -1$ for each pixel $i$.
Set distance $d(i) = \infty$ for each pixel $i$.
repeat
  /* Assignment */
  for each cluster center $C_k$ do
    for each pixel $i$ in a $2S \times 2S$ region around $C_k$ do
      Compute the distance $D$ between $C_k$ and $i$.
      if $D < d(i)$ then
        set $d(i) = D$
        set $r(i) = k$
      end if
    end for
  end for
  /* Update */
  Compute new cluster centers.
  Compute residual error $E$.
Until $E \leq threshold$
Output the segmented image $I$.
```

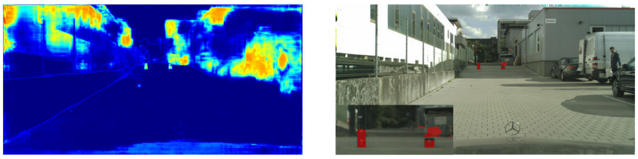

FIGURE 5. Anomaly scoring results: (a) Anomaly prediction map; (b) Anomaly score map.

region containing anomaly objects. Inspired by the obstacle scoring method in [37], the anomaly score calculation method above was designed in this paper to focus more on anomaly objects in the road. By assigning higher scores to these anomaly objects, more reasonable anomaly scoring results can be obtained, as shown in Figure 5. The anomaly scoring calculation method is shown in Algorithm 2.

IV. EXPERIMENTS AND SYSTEM EVALUATION

A. EXPERIMENTAL ENVIRONMENT AND PARAMETER SETTINGS

The experiment in this paper is based on the deep learning framework Pytorch, and the programming language is python3.8, the operating system is Windows11. In terms of experimental hardware, the CPU is Intel(R) Core(TM) i9-12900KF CPU @ 3.19GHz, the memory space is 128GB, the GPU is Nvidia Geforce RTX 3090, the video memory is 24GB, and the CUDA version is 11.3. The final anomaly score threshold was set at 0.5. During training, the Batch Size is 8, and the initial learning rate of the network is set to 0.0001. If there is no change in a certain index for more than 10 rounds, the learning rate is reduced. The Total Iteration is 50, and the network optimization method uses Adam, the momentum is 0.9, and the Cross Entropy Loss is used. During the training process and comparative experiments, the pre-trained model trained by the processed Lost and Found was used, and the mean and standard deviation values of ImageNet were used to normalize the training images to ensure the consistency of the experiment. And the training images are augmented by flipping around the vertical axis.

Algorithm 2 Anomaly Score Calculation

Input: *img*: Original input image; *superpixel*: Superpixel blocks after image segmentation; S_i: Anomaly calculation score

Output:*superpixels*: Array of superpixel blocks; *anomaly*_score: Anomaly score value for each superpixel block; *anomaly*_scores: Array of anomaly scoring results for each superpixel block

```
Segment the img with superpixels, and return the segmented array of superpixel blocks.
superpixels = perform_superpixel_segmentation(img)
for each superpixelinsuperpixels do
  ifsuperpixel contains anomaly objects then
    anomaly_score = calculate_anomaly_score($S_{i\_h}$)
  else
    anomaly_score = calculate_anomaly_score($S_{i\_l}$)
  end if
    Add anomaly_score to anomaly_scores array.
end for
return anomaly_scores
```

B. DATASETS AND PREPROCESSING

The dissimilarity module was trained using the Cityscapes dataset [38], evaluated by the FS Lost and Found dataset [39] from the Fishyscapes dataset, and used the Lost and Found dataset [40] and Obstacle Track dataset [41] as the testing dataset. As mentioned in the introduction, all objects that do not belong to the training class can be regarded as anomaly classes. Before training, the Cityscapes dataset needs to be preprocessed to obtain the original semantic map, synthetic map, Softmax distance map, and anomaly map, which are then jointly input to the dissimilarity module with the original input map and deep semantic images for training. In this paper, all void classes in the Cityscapes dataset are marked as 255 as anomalies. This method can cover any object that does not belong to the training class, which solves the problem of insufficient training data coverage scenarios. The area covered by void belongs to the anomaly area, which can be matched with high-uncertainty pixels, thereby guiding the

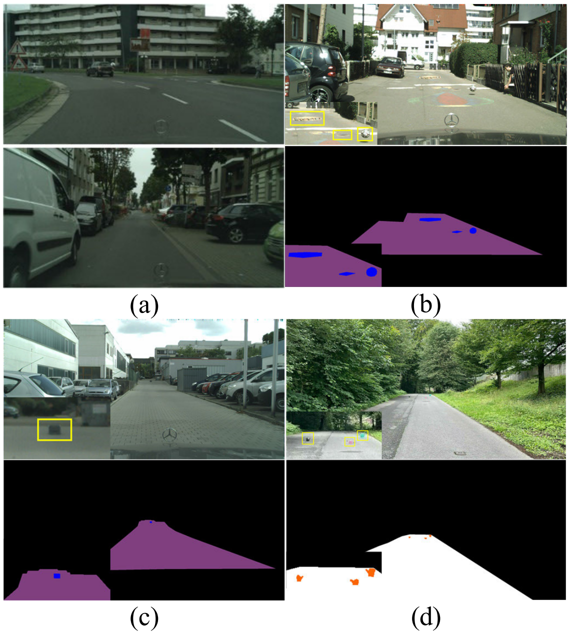

FIGURE 6. Dataset example: (a) Cityscapes; (b) FS lost and found; (c) Lost and found; (d) Obstacle track.

dissimilarity network to make full use of uncertainty information, so as to train a more robust anomaly detection model, but at the same time lose the advantage of not requiring OoD data during training. Nevertheless, the model still has good generalization and robustness to the detection of anomaly objects.

The FS Lost and Found dataset contains about 100 scene images taken from different streets with pixel-level semantic annotations of roads and anomaly objects. The Lost and Found dataset contains about 1023 images from various street scenes with pixel-level semantic annotations. The Obstacle Track dataset has 442 obstacle images with the road as the region of interest, and has pixel-level semantic annotations of anomaly, not anomaly and void classes (neither not anomaly nor anomaly). The Cityscapes dataset has 5950 images of driving scenes taken from different cities, which contains numerous categories, and provides corresponding semantic annotations. Examples of the four datasets are shown in Figure 6.

C. EVALUATION CRITERIA

This paper use pixel-level criteria to evaluation, where *AP* (Average Precision) is the pixel-level detection rate, the specific formula is as follows:

$$AP = \sum_{i=1}^{n-1} (r_{i+1} - r_i) \max_{r' \geq r_{i+1}} \rho(r') \tag{7}$$

Among them, r_i is the recall rate value corresponding to the first interpolation point of the precision rate interpolation section, and $\rho(r')$ represents the precision rate value when the recall rate is r'.

FPR (false positives rate) is the pixel-level false positive rate, the specific formula is as follows:

$$FPR = \frac{FP}{FP + TN} \tag{8}$$

Among them, *FP* represents the number of pixels wrongly detected as anomaly objects, and *TN* represents the number of pixels predicted as nonanomaly objects.

D. ANALYSIS OF EXPERIMENTAL RESULTS

The road anomaly objects detection framework proposed in this paper focuses on detecting various anomaly objects that appear within the road range. We use Cityscapes, a conventional road object classification scene dataset, as the basic training set, and the Lost and Found dataset and Obstacle Track dataset as the test set, which fully meets the purpose we want to achieve. The environmental scene categories we used for the test set are the same or similar to the training set, but the test set contains a large number of anomalous object categories not seen in the training set, so that we can use them to evaluate our method. The main body of the method in this paper is the anomaly detection module in part A of the third section. The dissimilarity network learns pixel-to-pixel comparison features during training, the model considers all pixel categories of the input image, so that we can Output an anomaly prediction map at the pixel level, as shown in Figure 4(a). But considering our task goals and the labels of Ground Truth in the test dataset, we only evaluate the pixel scene within the road range without considering pixels outside the road, which is consistent with the driving scene. Because in autonomous driving or driver assistance systems, the primary objective is to safely operate the vehicle, focusing attention on detecting anomaly objects within the road area directly serves the task goal and contributes to enhancing system performance. Additionally, concentrating attention within the road area reduces the complexity of information processing, leading to a reduction in computational and decision-making burdens, thereby facilitating more efficient execution of driving tasks. Therefore, after the anomaly prediction map is processed by the postprocessor, it will focus more on the situation within the road range. We also uniformly evaluate the effectiveness of different methods in detecting anomaly objects (i.e. OoD objects) within the road range to verify the effectiveness of our method. In our opinion, this is logically consistent with the driving scenario.

In order to evaluate the detection effect of the detection method MMF (Multiple Mechanism Fusion) in this paper on road anomaly objects, the method in this paper is compared with the following seven methods: 1) DeeplabV3+ [6]: it is a method that utilizes the encoding-decoding structure and the dilated Convolution to enhance the feature extraction effect on objects of different sizes; 2) Resynthesis [27]: This is a method of detecting anomalies by comparing the differences between the input image, the semantic segmentation image, and the generated image through the dissimilarity network; 3) Softmax Entropy [17]: It is a baseline method that

TABLE 1. Performance metrics comparison of the eight detection methods.

Method	Obstacle Track			L&F		
	AP↑	FPR95↓	mean F1↑	AP↑	FPR95↓	mean F1↑
DeeplabV3+ [6]	4.88	50.31	2.51	5.74	35.55	1.97
Resynthesis [27]	27.17	14.7	9.49	29.08	18.82	14.17
Resynthesis(w/ void)	28.29	15.24	8.55	31.11	16.03	12.84
Softmax Entropy [17]	55.41	5.28	11.06	68.34	9.53	18.43
Road Inpainting [30]	63.87	17.61	35.7	**82.63**	20.41	50.25
JSRNet [26]	60.53	**2.9**	13.82	71.41	**3.57**	36.17
DaCUP[31]	67.48	5.68	30.22	76.56	5.61	45.32
FlowEneDet[32]	70.06	4.79	37.48	79.13	5.69	48.31
MMF	**70.93**	4.16	**39.64**	80.04	5.31	**50.77**

measures uncertainty from the predicted softmax distribution and classifies samples as out-of-distribution samples through simple statistics. 4) Road Inpainting [30]: In this approach, the road drivable areas are first patched entirely, and then a dissimilarity network is utilized to identify the discrepancies between the original image and the patched image, thereby indicating the presence of anomaly objects that have been erased by the patches. 5) JSRNet [26]: This method employs a reconstruction module to identify and reconstruct road surfaces. The reconstruction module generates reconstruction errors, which are then coupled with semantic segmentation using trainable coupling blocks. This integration combines information from known classes and generates the final per-pixel anomaly scores for anomaly detection. 6) DaCUP [31]: This paper presents DaCUP, an anomaly detection method for autonomous driving. By employing a unique autoencoder-like architecture, image-conditioned distance features, and an inpainting module, and achieves well performance. However, it has limitations in detecting small and distant objects. 7) FlowEneDet [32]: This paper adopts a normal flow framework for improving the robustness of semantic segmentation models in real-world data settings with distribution shifts and outlier classes. It performs intradistribution misclassification (IDM) and outlier category detection simultaneously, and then implements a low-complexity 2D architecture through energy input, without the need for cumbersome retraining of the pre-trained semantic segmentation model.

Table 1 shows the performance metrics of MMF and the five above methods on Lost and Found and Obstacle Track datasets. During evaluation, the road area was partitioned into the region of interest (ROI), hence only the evaluation results within the ROI were considered, which aligns with normal driving routes. The results in the table demonstrate that MMF exhibits the best comprehensive performance among all six methods. Notably, only MMF and DeeplabV3+ utilized additional OoD data during training. However, MMF still exhibits strong generalization. Firstly, it can be observed that DeeplabV3+ performs poorly on both datasets, with very low AP scores and high false positives. During DeeplabV3+ training, an additional class was added to the Cityscapes dataset as an anomaly class, which participated in the training of this method. The evaluation results indicate that this approach is ineffective. Resynthesis is an earlier method that uses the difference between synthesized and original image feature to detect anomalies. The evaluation results demonstrate that this approach is effective for detecting anomaly objects. In the early framework, a random class is used to replace a randomly selected instance class from the ground truth semantic map in this method, then synthesized a new image using the replaced semantic map to compare the difference between the synthesized and original images. However, this approach has poor robustness and cannot cover too many anomaly situations, resulting in poor performance metrics. Resynthesis method was trained using the Cityscapes dataset, which contains OoD data. The results indicate that the performance improvement is limited. Softmax Entropy performed well on both datasets, but its performance metrics were still poorer than MMF. This is because single uncertainty estimates cannot solve boundary problems, resulting in a high false positive rate. The Road Inpainting method also employs the generation of new images and utilizes a dissimilarity network to detect anomalies by comparing the features between the original and reconstructed images. The evaluation results indicate that this method achieves relatively high AP scores on both datasets, particularly outperforming others on the Lost and Found dataset. However, it is expected that this method exhibits high false positive rates on both datasets. This is not surprising since similar to Resynthesis, these methods lack constraints when detecting differences between the original and reconstructed images, making them more susceptible to noise interference and resulting in a higher number of false positives. JSRNet demonstrates good performance on both datasets, with relatively low false positive rates. However, it ranks fifth in terms of AP among the six methods. This method employs a network similar to an auto-encoder to learn the discriminative reconstruction of RGB values for road pixels. The major drawback of auto-encoder methods is the simplistic bottleneck that fails to effectively utilize the

TABLE 2. Abalation experiments.

Method	Obstacle Track		L&F	
	AP↑	**FPR95↓**	**AP↑**	**FPR95↓**
w/o ass. maps	53.13	4.19	56.02	6.36
w/o postprocessor	66.88	13.94	74.41	13.83
w/o ass. maps & w/o postprocessor	28.29	15.24	31.11	16.03
Full Framework	**70.93**	**4.16**	**80.04**	**5.31**

training data. This limitation hinders the network's ability to learn relevant features, resulting in decreased accuracy. DaCUP performed well in the AP index in the two data sets, ranking third among all comparison methods, but the false positive rate was high. This method does not use lateral skip connected low-resolution features in the upsampling process. This may cause the model to have difficulty in recovering the details of small targets, thereby affecting detection performance and leading to an increase in the false positive rate. FlowEneDet performs equally well on both data sets, with high accuracy and ranks second in AP index among all methods, but its false positive rate is high. This paper addresses the challenges of distribution shifts and outlier classes, where models may be more prone to false positives. Distribution shifts may cause the model to fail to accurately generalize to new data distributions when tested, thereby increasing misclassification of normal samples. The mean F1 scores for each comparative method were also analyzed, revealing that MMF consistently achieves the highest mean F1 scores on both datasets. This suggests that the MMF method strikes a good balance in anomaly detection tasks and accurately identifies anomaly objects. Overall, MMF demonstrates the best comprehensive performance on both datasets. Given that the objective of this paper is to iden-tify road anomalies, the goal of our work is to obtain conservative results, meaning higher recognition accuracy w-hile maintaining a low false positive rate. So, MMF has the best performance metrics on both datasets, primarily because the assistant methods proposed in MMF complement the dissimilarity network very well. The uncertainty estimates and depth information provided features that complement the dissimilarity network, enhancing the network's ability to locate anomaly objects and improving detection performance. Furthermore, the postprocessor method further reduces false positives generated by network predictions, enhancing the robustness of the dissimilarity network. Compared with detection methods that do not use OoD data, the method in this paper not only improves the detection rate but also maintains good generalization.

Table 2 shows the ablation experimental results of the method in this paper, which shows the influence and contribution of each module in the MMF method to the overall method. The results show that the framework performs extremely poorly on both datasets without the assistant maps and postprocessor, with AP metrics of only 28.29% and 31.11%, and FPR95 metrics of 15.24% and 16.03%, respectively. Such low detection rates and false positive rates can be fatal in practical tasks. The addition of assistant maps and postprocessor proposed in MMF has significantly improved the framework. The AP metrics of assistant maps on the two datasets have been greatly improved. This also proves that the feature information between assistant maps and resynthesis methods is complementary, so adding assistant maps can make the framework achieve better performance. At the same time, it can be seen that adding assistant maps does not significantly improve the FPR95 index. This is expected, as it is a drawback of the uncertainty and depth maps in assistant maps themselves, and the dissimilarity network will learn more irrelevant feature differences. The assistant maps did not help the improvement of false positives very well.

Therefore, postprocessor is proposed in the MMF method to solve this problem. The results of the postprocessor on both datasets show a considerable improvement in the FPR95 metric, whereas its contribution to the AP metric is lower than that of the assistant maps. This is attributed to the anomaly scoring mechanism of the postprocessor. Postprocessor segments the predicted scenes into superpixel blocks and assigns an anomaly score to each block to filter the anomalous objects based on the set threshold. However, some irrelevant features learned in the dissimilarity network tend to get relatively low anomaly scores, so it is easy to ignore these irrelevant regions with low scores, thereby reducing false positives. As shown in Table 2, when all modules are integrated into the framework, MMF obtains higher detection rate and lower false positive rate, which is in line with expectations. Both assistant maps and postprocessor make contributions to the framework. Assistant maps focus more on enhancing the detection ability of the framework, while postprocessor focuses more on reducing the false positive rate of the framework. The results for individual modules and for all modules combined are given in the table.

Table 3 shows the inference times of each module in our framework, along with a comparison to the Resynthesis [27] method. The metrics were the average results after running 100 times on an NVIDIA RTX3090 with 24GB of memory. The input resolutions for each module are provided in the tabl- e. The results in the table showcase that the modules in our approach have improved the overall inference speed of the framework, reducing computational costs.

TABLE 3. Comparison on computational cost.

Module	MMF	Resynthesis	Resolution
Segmentation	686.93	875.68	2048×1024
Synthesis	105.00	133.86	1024×512
Depth	75.70	-	1024×512
Perceptual Difference	7.11	-	512×256
Dissimilarity	28.98	36.95	512×256
Total(ms)	903.72	1046.49	-

TABLE 4. Performance comparison between best and lighter frameworks (✔ This means the use of a lighter network. ✘ This means the use of a better network).

segmentation	synthesis	Lost and Found	
		AP↑	FPR95↓
✔	✘	61.38	10.06
✘	✔	75.64	6.92
✔	✔	57.63	10.17
✘	✘	80.04	5.31
Resynthesis [27]		29.08	18.82

In order to validate the generalization capability of the dissimilarity network approach proposed in this paper, we further conducted experiments by selecting different segmentation and synthesis networks than those mentioned in Section III. This paper chose Enet [14] as the segmentation network and SPADE [34] as the synthesis network. This combination formed a lighter network framework but with lower segmentation and synthesis performance. Consequently, we performed ablation experiments on these two newly selected components. The results of the ablation experiments are presented in Table 4. From the results, it can be observed that the performance of the dissimilarity network is influenced by both the segmentation and synthesis networks, with the segmentation network's quality being the primary factor affecting the final results. Regardless of the choice of segmentation and synthesis networks, the proposed method in this paper outperforms the original Resynthesis method by a significant margin. The experimental results also indicate that the better the performance of these two modules, the easier it becomes for the dissimilarity network to differentiate the feature differences between the original and synthesized images, which aligns with previous experience.

Figures 7 and 8 show three scenes from the Obstacle Track and Lost and Found datasets, respectively, using proposed MMF method. And the qualitative comparison between the method, uncertainty estimation method, and Resynthesis method is also shown. Selected images include shaded scenes, distant objects, and instances with unusual objects. As shown in Figure 7, the objects in the Obstacle Track dataset are small and flat, and MMF outperforms other methods in all three different scenes. It can be seen that the traditional semantic segmentation method is almost unavailable in such tasks, and that anomaly objects detection is basically wrong or not detected at all. Uncertainty estimation performs better in detecting nonflat anomaly objects than flat ones, but it still misses some of them, and its detection has weak generalization. The same problem is also encountered by the Resynthesis method. Its generalization ability for anomaly object detection is very weak, and many false detection areas are generated, resulting in inaccurate localization of anomaly objects. As shown in Figure 8, the scenes and anomaly objects in the Lost and Found dataset are similar to those in the Obstacle Track dataset. In three different scenarios, the segmentation network is still unable to complete the detection of such tasks. The uncertainty estimation method performs better for detecting regular-shaped anomalies such as boxes compared to other types of anomalies. However, it can be observed in the figure that many areas are still missed by the model. The false detection rate of the Resynthesis method in the Lost and Found dataset is still high, which further verifies that the original framework is extremely susceptible to noise interference. At this time, the

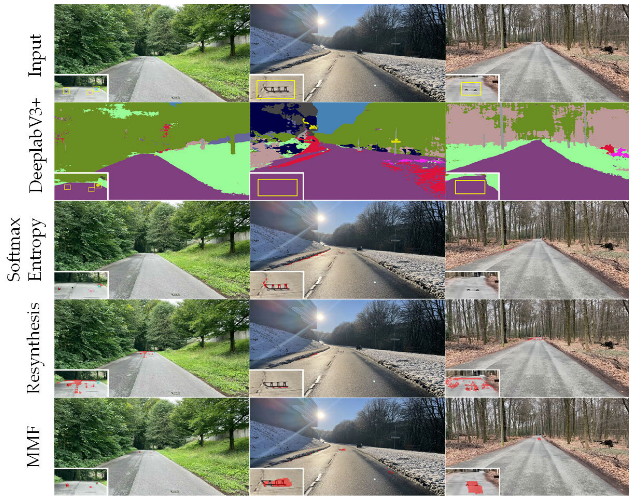

FIGURE 7. Obstacle track dataset detection results.

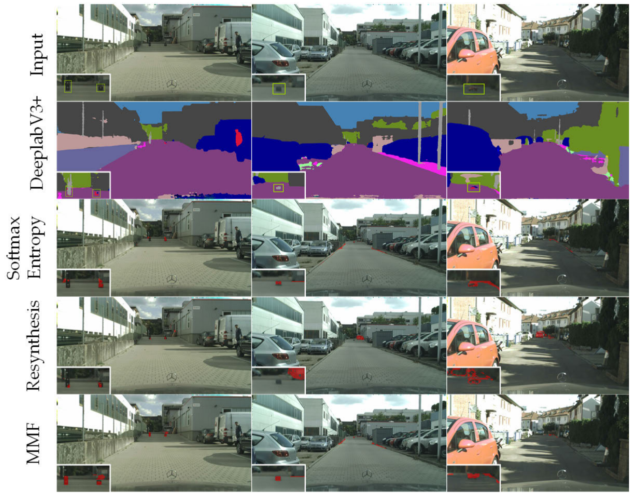

FIGURE 8. Lost and found dataset detection results.

detection ability of the dissimilarity network is at a low level. The MMF method, which integrates uncertainty mapping and depth mapping with the Resynthesis method, enhances the detection performance of the dissimilarity network on both datasets. It achieves high detection rates and precise localization of anomaly objects, and is further optimized by a postprocessor that reduces false positive rate and focuses on anomaly objects on the road. Notably, the MMF method has not encountered any anomaly objects used in testing during training, and the class marked as void is not within the visualization range.

V. CONCLUSION

In order to achieve accurate of anomaly objects on the road, a novel anomaly detection method is proposed in this paper. The method includes two parts: an anomaly detection framework and a postprocessor. The anomaly detection framework combines two methods that can complement each other with the resynthesis method, namely an uncertainty estimation method and a segmentation method that incorporates depth information. In the uncertainty estimation method, a new anomaly map is obtained by multiplying Softmax entropy and perceptual loss. It is then used together with the Softmax distance and depth maps as the attention of the dissimilarity network to guide it to focus on the feature differences between the input image and the generated image. Finally, the postprocessor is used to process the results of the anomaly detection framework, and the final prediction result is obtained. The experimental results on the Obstacle Track dataset and the Lost and Found dataset show that the proposed method achieves detection accuracies of 70.93% and 80.04%, with false positive rates of 4.16% and 5.31%, respectively. Its performance is much better than other methods in this paper. There is still room for improvement in terms of detection accuracy and false positive rate, even though the postprocessor has reduced the latter to a certain extent. Therefore, our subsequent work will continue to investigate how to educe the false positive rate of the prediction results to an ideal state, as well as explore how to achieve existing or better performance for the framework without relying on OoD data.

REFERENCES

[1] G. Zhang, K. K. W. Yau, X. Zhang, and Y. Li, "Traffic accidents involving fatigue driving and their extent of casualties," *Accident Anal. Prevention*, vol. 87, pp. 34–42, Feb. 2016.

[2] O. Gietelink, J. Ploeg, B. De Schutter, and M. Verhaegen, "Development of advanced driver assistance systems with vehicle hardware-in-the-loop simulations," *Vehicle Syst. Dyn.*, vol. 44, no. 7, pp. 569–590, Jul. 2006.

[3] K. Chu, M. Lee, and M. Sunwoo, "Local path planning for off-road autonomous driving with avoidance of static obstacles," *IEEE Trans. Intell. Transp. Syst.*, vol. 13, no. 4, pp. 1599–1616, Dec. 2012.

[4] C.-J. Li, Z. Qu, S.-Y. Wang, and L. Liu, "A method of cross-layer fusion multi-object detection and recognition based on improved faster R-CNN model in complex traffic environment," *Pattern Recognit. Lett.*, vol. 145, pp. 127–134, May 2021.

[5] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 6230–6239.

[6] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder–decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis.*, Munich, Germany, Sep. 2018, pp. 801–818.

[7] A. Bhattacharyya, M. Fritz, and B. Schiele, "Long-term on-board prediction of people in traffic scenes under uncertainty," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 4194–4202.

[8] A. Kendall and Y. Gal, "What uncertainties do we need in Bayesian deep learning for computer vision?" in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, Dec. 2017, pp. 5580–5590.

[9] 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 Robot. Autom. Lett.*, vol. 5, no. 4, pp. 5558–5565, Oct. 2020.

[10] Y. Li, C. Jung, and J. Kim, "Single image depth estimation using edge extraction network and dark channel prior," *IEEE Access*, vol. 9, pp. 112454–112465, 2021.

[11] G. Lv, S. M. Israr, and S. Qi, "Multi-style unsupervised image synthesis using generative adversarial nets," *IEEE Access*, vol. 9, pp. 86025–86036, 2021.

[12] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, Jun. 2015, pp. 3431–3440.

[13] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder–decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.

[14] A. Paszke, A. Chaurasia, S. Kim, and E. Culurciello, "ENet: A deep neural network architecture for real-time semantic segmentation," 2016, *arXiv:1606.02147*.

[15] J. E. Valdez-Rodríguez, H. Calvo, E. Felipe-Riverón, and M. A. Moreno-Armendáriz, "Improving depth estimation by embedding semantic segmentation: A hybrid CNN model," *Sensors*, vol. 22, no. 4, p. 1669, Feb. 2022.

[16] H. Wang, R. Fan, Y. Sun, and M. Liu, "Dynamic fusion module evolves drivable area and road anomaly detection: A benchmark and algorithms," *IEEE Trans. Cybern.*, vol. 52, no. 10, pp. 10750–10760, Oct. 2022.

[17] D. Hendrycks and K. Gimpel, "A baseline for detecting misclassified and out-of-distribution examples in neural networks," in *Proc. Int. Conf. Learn. Represent.*, Toulon, France, Apr. 2017, pp. 24–26.

[18] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, pp. 1929–1958, Jun. 2014.

[19] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in *Proc. Int. Conf. Mach. Learn.*, New York, NY, USA, Jun. 2016, pp. 1050–1059.

[20] A. Kendall, V. Badrinarayanan, and R. Cipolla, "Bayesian SegNet: Model uncertainty in deep convolutional encoder–decoder architectures for scene understanding," in *Proc. Brit. Mach. Vis. Conf.*, 2017, p. 57.

[21] S. Isobe and S. Arai, "A semantic segmentation method using model uncertainty," in *Proc. 5th IIAE Int. Conf. Ind. Appl. Eng.*, Kitakyushu, Japan, 2017, pp. 89–93.

[22] J. Mukhoti and Y. Gal, "Evaluating Bayesian deep learning methods for semantic segmentation," 2018, *arXiv:1811.12709*.

[23] M. Rottmann, P. Colling, T. P. Hack, R. Chan, F. Hüger, P. Schlicht, and H. Gottschalk, "Prediction error meta classification in semantic segmentation: Detection via aggregated dispersion measures of softmax probabilities," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2020, pp. 1–9.

[24] P. Oberdiek, M. Rottmann, and G. A. Fink, "Detection and retrieval of out-of-distribution objects in semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Seattle, WA, USA, Jun. 2020, pp. 1331–1340.

[25] C. Baur, B. Wiestler, S. Albarqouni, and N. Navab, "Deep autoencoding models for unsupervised anomaly segmentation in brain MR images," in *Proc. Int. MICCAI Brainlesion Workshop*, Granada, Spain, Sep. 2018, pp. 161–169.

[26] 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 *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Montreal, QC, Canada, Oct. 2021, pp. 15631–15640.

[27] K. Lis, K. K. Nakka, P. Fua, and M. Salzmann, ''Detecting the unexpected via image resynthesis,'' in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 2152–2161.
[28] Y. Xia, Y. Zhang, F. Liu, W. Shen, and A. L. Yuille, ''Synthesize then compare: Detecting failures and anomalies for semantic segmentation,'' in *Proc. Eur. Conf. Comput. Vis.*, Glasgow, U.K., Aug. 2020, pp. 145–161.
[29] G. Di Biase, H. Blum, R. Siegwart, and C. Cadena, ''Pixel-wise anomaly detection in complex driving scenes,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Nashville, TN, USA, Jun. 2021, pp. 16913–16922.
[30] K. Lis, S. Honari, P. Fua, and M. Salzmann, ''Detecting road obstacles by erasing them,'' 2020, *arXiv:2012.13633*.
[31] T. Vojír and J. Matas, ''Image-consistent detection of road anomalies as unpredictable patches,'' in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Waikoloa, HI, USA, Jan. 2023, pp. 5480–5489.
[32] D. Gudovskiy, T. Okuno, and Y. Nakata, ''Concurrent misclassification and Out-of-Distribution detection for semantic segmentation via energy-based normalizing flow,'' 2023, *arXiv:2305.09610*.
[33] X. Liu, G. Yin, J. Shao, and X. Wang, ''Learning to predict layout-to-image conditional convolutions for semantic image synthesis,'' in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32. Vancouver, BC, Canada, Dec. 2019, pp. 570–580.
[34] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu, ''Semantic image synthesis with spatially-adaptive normalization,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 2332–2341.
[35] R. Kamoi, T. Iida, and K. Tomite, ''Efficient unknown object detection with discrepancy networks for semantic segmentation,'' in *Proc. NeurIPS Workshop Mach. Learn. Auton. Driving*, Dec. 2021, pp. 1–12.
[36] K. He, X. Zhang, S. Ren, and J. Sun, ''Identity mappings in deep residual networks,'' in *Proc. Eur. Conf. Comput. Vis.*, Amsterdam, The Netherlands, Oct. 2016, pp. 630–645.
[37] T. Ohgushi, K. Horiguchi, and M. Yamanaka, ''Road obstacle detection method based on an autoencoder with semantic segmentation,'' in *Proc. Asian Conf. Comput. Vis.*, Kyoto, Japan, Nov. 2020, pp. 223–238.
[38] 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 *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 3213–3223.
[39] H. Blum, P.-E. Sarlin, J. Nieto, R. Siegwart, and C. Cadena, ''The fishyscapes benchmark: Measuring blind spots in semantic segmentation,'' *Int. J. Comput. Vis.*, vol. 129, no. 11, pp. 3119–3135, Nov. 2021.
[40] P. Pinggera, S. Ramos, S. Gehrig, U. Franke, C. Rother, and R. Mester, ''Lost and found: Detecting small road hazards for self-driving vehicles,'' in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2016, pp. 1099–1106.
[41] R. Chan, K. Lis, S. Uhlemeyer, H. Blum, S. Honari, R. Siegwart, P. Fua, M. Salzmann, and M. Rottmann, ''SegmentMeIfYouCan: A benchmark for anomaly segmentation,'' 2021, *arXiv:2104.14812*.

WENYAN CI received the M.S. degree from Nanjing Normal University, China, in 2010, and the Ph.D. degree from the University of Shanghai for Science and Technology, China, in 2019. He is currently an Assistant Professor with Huzhou University, Zhejiang, China. He has undertaken several provincial and municipal projects. He has published many influential articles in the field of intelligent vehicle perception technology. His research interests include intelligent vehicle perception technology based on machine vision, including intelligent vehicle obstacle detection and recognition, visual odometry, and simultaneous localization and mapping (SLAM).

JIAYIN XUAN received the B.S. degree in automation from the Wanjiang College, Anhui Normal University, Wuhu, China, in 2020. He is currently pursuing the M.S. degree in electronic information with Huzhou University, Zhejiang, China. His research interests include computer vision, intelligent vehicles, and environmental perception.

RUNZE LIN received the B.S. degree in automation from the College of Control Science and Engineering, Zhejiang University, Hangzhou, China, in 2020, where he is currently pursuing the Ph.D. degree with the State Key Laboratory of Industrial Control Technology, China. He was a Visiting Scholar with the University of Alberta, Edmonton, Canada, from 2022 to 2023. His research interests include reinforcement learning, transfer learning, data analytics, and industrial big data and its applications.

SHAN LU received the B.S. and Ph.D. degrees from Zhejiang University, China, in 2011 and 2016, respectively. He is currently an Assistant Professor and the Deputy Director with the Institute of Intelligence Science and Engineering, Shenzhen Polytechnic. He has undertaken several national and provincial funds and projects in the field of smart manufacturing and process control systems. He acts as the Principal Investigator of Innovation Team of Guangdong Province, China. His research interests include data-driven optimization, cyber system control, and signal processing.