InconSeg: Residual-Guided Fusion With Inconsistent Multi-Modal Data for Negative and Positive Road Obstacles Segmentation

Zhen Feng, Yanning Guo, David Navarro-Alarcon, Yueyong Lyu, and Yuxiang Sun

***Abstract*—Segmentation of road obstacles, including negative and positive obstacles, is critical to the safe navigation of autonomous vehicles. Recent methods have witnessed an increasing interest in using multi-modal data fusion (e.g., RGB and depth/disparity images). Although improved segmentation accuracy has been achieved by these methods, we still find that their performance could be easily degraded if the two modalities have inconsistent information, for example, distant obstacles that can be viewed in RGB images but cannot be viewed in depth/disparity images. To address this issue, we propose a novel two-encoder-two-decoder RGB-depth/disparity multi-modal network with Residual-Guided Fusion modules. Different from most existing networks that fuse feature maps in encoders, we fuse feature maps in decoder. We also release a large-scale RGB-depth/disparity dataset recorded in both urban and rural environments with manually-labeled ground truth for both negative- and positive-obstacles segmentation. Extensive experimental results demonstrate that our network achieves state-of-the-art performance compared with other networks.**

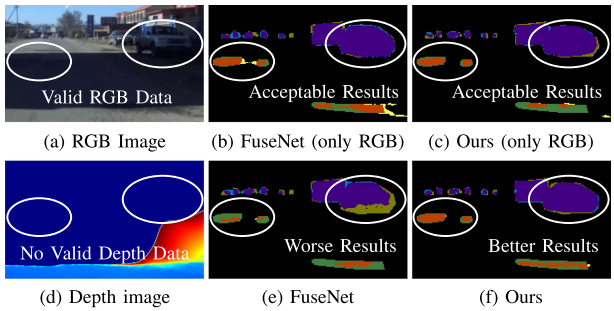
***Index Terms*—Negative obstacles, road obstacles, multi-modal fusion, semantic segmentation, autonomous vehicles.**

I. Introduction

ROAD-OBSTACLE segmentation is a fundamental capability for autonomous vehicles. Road obstacles can be generally divided into two categories: positive obstacles and negative obstacles. Positive obstacles refer to those that stand on the ground, such as pedestrians, vehicles, bicycles, etc. These obstacles have attracted great attention in the robotics research community because detection or segmentation of positive obstacles is critical to downstream tasks, such as tracking [1] and navigation [2], etc. Negative obstacles refer to those that have heights lower than the ground, such as potholes, cracks, etc. They can cause discomfort to passengers. Large negative obstacles can even cause accidents, such as roll over to vehicles [3]. So, accurate detection or segmentation of negative obstacles is also of great importance to the safety of autonomous vehicles.

Many effective segmentation methods have been respectively proposed for positive obstacles [4] and negative obstacles [5]. But most of them use single-modal visual data (e.g., only RGB images). Since visual images could be affected by environmental lighting conditions, these methods could be easily degraded. To address this problem, many multi-modal (e.g., RGB-depth/disparity, RGB-thermal) fusion networks have been proposed. These networks could produce superior results over single-modal networks [6], [7], [8].

However, we observe that when there exists inconsistency between different modals of data, the segmentation performance of these data-fusion networks is not better or even inferior to those using a single modal of data. Note that in this letter, inconsistency refers to that one modal contains some information, but the other modal does not have the corresponding information in the same region. Fig. 1 shows an example for such inconsistency. We can see that the vehicle highlighted in the white ovals is too far (out of the depth measurement range), so that the depth data become invalid. In addition, due to shadows, there is no valid depth data at the potholes. Our experiments show that data-fusion



(a) RGB Image (b) FuseNet (only RGB) (c) Ours (only RGB)

(d) Depth image (e) FuseNet (f) Ours

Fig. 1. Negative effects caused by the inconsistency between the RGB-depth images. In this figure, depth increases from red to blue. The white oval in (a) and (d) highlight the same area, where the inconsistency can be clearly seen, that is, there is no valid information in the depth image but there is valid information in the RGB image. In (b) and (c), 'only RGB' means the network is trained and tested with only RGB images. We can see that the inconsistency leads to inferior segmentation performance in these areas for FuseNet, but our network can give better results. ■, ■, and ■ refer true positive, false negative, and false positive of negative obstacles. ■, ■, and ■ refer true positive, false negative, and false positive of positive obstacles.

Manuscript received 18 November 2022; accepted 3 April 2023. Date of publication 2 May 2023; date of current version 5 July 2023. This letter was recommended for publication by Associate Editor Y. Joo and Editor H. Moon upon evaluation of the reviewers' comments. This work was supported in part by the National Natural Science Foundation of China under Grant 62003286, in part by Zhejiang Lab under Grant 2021NL0AB01, in part by the Guangdong Basic and Applied Basic Research Foundation under Grant 2022A1515010116, in part by CCF-Baidu Open Fund under Grant 182215PCK04183, and in part by the Start-up Fund of HK PolyU under Grant P0034801. *(Corresponding author: Yuxiang Sun.)*

Zhen Feng is with the Department of Mechanical Engineering, The Hong Kong Polytechnic University, Hong Kong, China, and also with the Department of Control Science and Engineering, Harbin Institute of Technology, Harbin 150001, China (e-mail: zfeng94@outlook.com).

Yanning Guo and Yueyong Lyu are with the Department of Control Science and Engineering, Harbin Institute of Technology, Harbin 150001, China (e-mail: guoyn@hit.edu.cn; lvyy@hit.edu.cn).

David Navarro-Alarcon and Yuxiang Sun are with the Department of Mechanical Engineering, The Hong Kong Polytechnic University, Hong Kong, China (e-mail: dnavar@polyu.edu.hk; yx.sun@polyu.edu.hk).

Digital Object Identifier 10.1109/LRA.2023.3272517

networks (e.g., FuseNet [7]) with inconsistent RGB-depth data present inferior segmentation performance to those with only RGB images.

To provide a solution to this issue, we propose a novel RGB-depth/disparity fusion network with Residual-Guided Fusion modules in a two-encoder-two-decoder structure. Our fusion strategy is expected to extract information from one modality that is missing from the other so that the information could be complemented, and hence increase the segmentation performance. Fig. 1 shows that our network produces better results than FuseNet when given two modalities with inconsistent information, and also shows that our network produces better results than a single modality when fusing inconsistent RGB-depth images.

We also note that for positive-obstacle segmentation, there exist a large number of datasets in the research community [9], but for negative obstacles, there exist limited datasets. Moreover, existing negative-obstacle datasets are collected mainly from urban environments. So, we build a large-scale dataset with manually labelled masks for negative and positive obstacles in both urban and rural environments. This dataset is captured by a ZED stereo camera with inconsistent RGB-depth/disparity (RGB-D) images. The main contributions of this letter are summarized as follows:

1) We propose a novel RGB-D fusion network for negative and positive road obstacles segmentation in a two-encoder-two-decoder structure.
2) We design a novel Residual-Guided Fusion (RGF) module to address the inconsistency issue between multi-modal data through extracting complementary features for the missing features of RGB images from depth images.
3) We release a large-scale RGB-D dataset (there are 5,000 images with manually-labelled ground truth) for segmentation of negative and positive road obstacles. Moreover, our code is open-sourced[1].

II. Related Work

A. Single-Modal Semantic Segmentation Networks

Chen et al. [10] introduced atrous convolution into spatial pyramid pooling to increase receptive fields, and designed DeepLabV3+ in the encoder-decoder structure [11]. Recently, Transformer has achieved superior performance in computer vision. Many semantic segmentation networks based on the transformer structure have been proposed. For example, Azad et al. [12] combined the transformer structure with DeepLabV3+ to design TransDeepLab for medical image segmentation.

B. Multi-Modal Semantic Segmentation Networks

Multi-modal fusion is commonly achieved by feature addition or concatenation. For example, Hazirbas et al. [7] proposed FuseNet with an RGB encoder and a depth encoder to fuse RGB-depth images by element-wise addition. Sun et al. [13] proposed RTFNet by fusing RGB and thermal images in the two-encoder-one-decoder structure by element-wise addition. Fan et al. [6] design AA-RTFNet by introducing attention modules into RTFNet as skip connections between the encoder and the decoder. Some networks fuse multi-modal features with attention modules. For example, Seichter et al. [8] designed an RGB-depth fusion layer with a squeeze and excitation module in ESANet. Sun et al. [14] proposed RFNet with an attention feature complementary module to fuse RGB-depth images for segmenting obstacles on roads. Feng et al. [3] proposed MAFNet that fuses RGB feature maps and disparity feature maps using channel attention modules and dual attention modules. Ying et al. [15] proposed UCTNet with an uncertainty-aware self-attention module to avoid the influence of unreliable information in depth images. Some networks specially design fusion modules to fuse multi-modal features. For example, Chen et al. [16] proposed SA-Gate module to fuse RGB-HHA images that firstly separates features from both modalities and then aggregates features to generate fusion results. Valada et al. [17] proposed the CMoDE fusion framework to adaptively fuse modalities for the alleviation of the influence caused by environmental condition changes. The CMoDE fuses the predicted maps of each modality via a domain-expert method. Valada et al. [18] proposed UpNet and adopted a late-fused convolution technique to fuse multi-modal data. They released an outdoor dataset with multi-spectral images to evaluate their network. Pfeuffer et al. [19] proposed Faster-LSTM-ICNet that speeds up LSTM-ICNet and achieves robust segmentation performance in adverse weather conditions. Wang et al. [20] proposed the transformer-based TokenFusion to learn the correlations among multi-modal features. TokenFusion aligns multi-modal features by residual positional alignment strategy after fusion.

C. Road-Obstacle Datasets

Pinggera et al. [21] released a dataset with 2,104 labelled frames for detecting small obstacles. The dataset is collected from 13 street scenarios. There are two classes (i.e., obstacle and free space) in the dataset. Fan et al. [6] released the Pothole-600 dataset with 600 pairs of RGB and transformed disparity images for road potholes segmentation. Han et al. [22] released the Puddle-1000 dataset with 985 images for water puddles segmentation. The authors captured images with a ZED camera. The ground truth masks are labeled on the left images. Bijelic et al. [23] released a multi-modal adverse weather dataset captured by a stereo camera for object detection, gated camera, radar, LiDAR, and far-infrared camera. They also proposed a real-time multi-modal fusion network to fuse these modalities.

D. Difference With Existing Works

Different from the aforementioned multi-modal fusion networks, our network adopts a two-encoder-two-decoder structure and fuses the feature maps restored by multiple stages of the decoder. We quantify missing features of the RGB modality through an RGF module. We also extract complementary features for the missing features from another modality through the RGF module. Moreover, different from the aforementioned datasets, our dataset contains more images with both negative and positive obstacles.

III. The Proposed Network

A. The Overall Architecture

The overall architecture of our proposed InconSeg is shown in Fig. 2. There are two data streams: a depth/disparity stream and an RGB stream. The depth/disparity stream takes as input depth or disparity images. Since depth and disparity images can be easily converted to each other, in the following text, we use

[1]Our code and dataset: https://github.com/lab-sun/InconSeg.

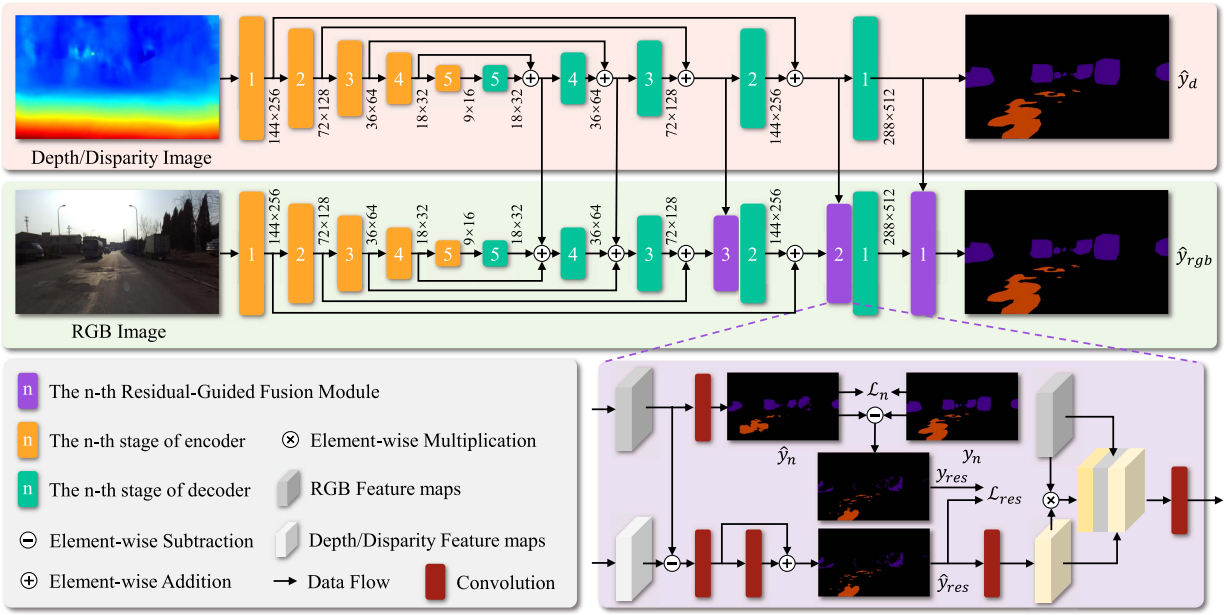


Fig. 2. Overall architecture of our proposed InconSeg. There are two streams: an RGB stream and a depth/disparity stream. Each stream has a 5-stage encoder and 5-stage decoder. The output of the RGB stream $\hat{y}_{rgb}$ is the output of InconSeg. The encoder is adopted from ResNet-152 [24]. The outputs of the first three stages of the depth/disparity decoder are fused into the stages of the RGB decoder at the same level by our proposed Residual-Guided Fusion (RGF) module.

depth instead of depth/disparity for convenience. Each stream has a 5-stage encoder and a 5-stage decoder. The output of the RGB stream $\hat{y}_{rgb}$ is the output of the network InconSeg. In the encoders, each stage reduces the resolution of the input image by half. In the decoders, each stage doubles the resolution and reduces the number of channels of feature maps by half. The output of the depth stream $\hat{y}_d$ is only used during training. The encoders are borrowed from ResNet-152 [24].

In each stream, the inputs of each stage of the encoder are fused with the output of each stage of the decoder at the same level by element-wise addition. Different from existing networks, such as MAFNet [3] and AA-RTFNet [6], we fuse the outputs of both decoders at the same level to avoid the negative effects caused by the inconsistency of the feature maps extracted by the encoders. The outputs of the last two stages of the depth decoder are fused into the same-level stage of the RGB decoder via element-wise addition. The outputs of the first three stages of the depth stream decoder are fused with the output of the same-level stage of the RGB decoder by our proposed RGF module. The three RGF modules are placed in the RGB decoder. The n-th RGF module is placed behind the n-th stage of the RGB decoder, where $n \in [1, 2, 3]$.

B. The RGF Module

As aforementioned, the purpose of the RGF module is to quantify the missing features between the RGB features and the ground truth. The RGF module extracts the complementary features of the RGB features from the depth features instead of directly fusing them, thus addressing the degradation of fusion performance caused by inconsistent data between both features. The structure of our RGF module is shown on the right bottom of Fig. 2. The module has two inputs: RGB feature maps and depth feature maps.

Firstly, the RGF module generates the missing features of RGB modality. Specifically, the RGB feature maps generate RGB predicted mask $\hat{y}_n$ through a 1×1 convolutional layer, where n represents the n-th RGF module. It should be noted that the first RGF module does not contain the 1×1 convolution. The residual mask y^n_{res} is generated through an element-wise subtraction between $\hat{y}_n$ and the ground truth y_n. The residual mask y^n_{res} represents the residual features. We call y^n_{res} as missing features of the RGB feature map. $\hat{y}_n$ and y_n have the same resolution as the RGB feature maps. y_n is generated from the original ground truth y with down-sampling using the nearest neighbor method.

Secondly, we extract complementary features for the missing features. Specifically, we subtract the RGB feature maps with depth feature maps by element-wise subtraction to get the difference between them. The channel of the different features is adjusted to the number of classes through a 1×1 convolution. It should be noted that the first RGF module does not contain the 1×1 convolution. Then, a residual unit with a 3×3 convolution is used to generate the predicted residual mask $\hat{y}^n_{res}$. y^n_{res} is used to guide the generation of $\hat{y}^n_{res}$. The channel of $\hat{y}^n_{res}$ is adjusted to that of the RGB feature maps by a 1×1 convolution. After that, the adjusted result is fused with the RGB feature maps through an element-wise multiplication. Finally, the adjusted result, fusion result, and RGB feature maps are concatenated along the channel dimension. The output of the RGF module is generated by a 1×1 convolution, which is fed into the next stage in the RGB decoder.

C. The Structure of Decoders

The structure of each stage of the decoders is shown in Fig. 3. Firstly, the input feature map is fed into a two-branch residual structure. The residual structure reduces the resolution of the input feature map by half. The upper branch has a 1×1 Convolution-BN-ReLU layer. The lower branch has three 3×3 Convolution-BN-ReLU layers. The first layer in the lower branch reduces the resolution of the input feature map by half. The other layers in the lower branch keep the number of channels unchanged. The outputs of both branches are fused through

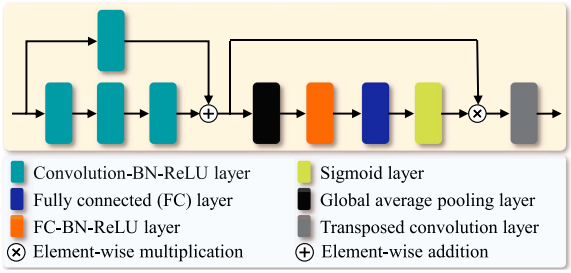


Fig. 3. Structure of each stage of the decoder. The input feature map first passes through a two-branch residual structure. The upper branch has one Convolution-BN-ReLU layer. The lower branch has three sequential Convolution-BN-ReLU layers.

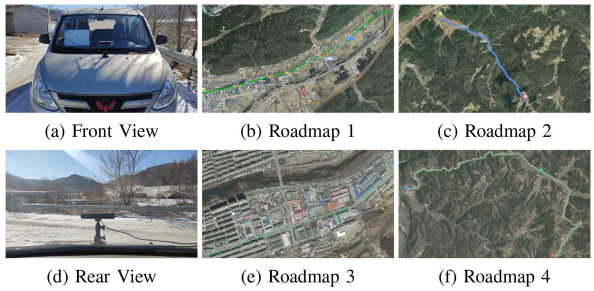


Fig. 4. Equipment for data collection and some sample roadmaps for data collection. Roadmap 1 and Roadmap 3 are urban scenes. Roadmap 2 and Roadmap 4 are rural scenes. The scenes are in Liaoning Province, China.

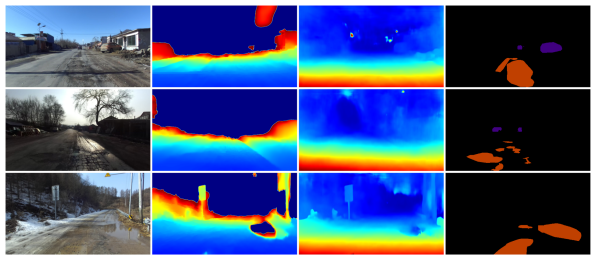


Fig. 5. Sample RGB images, depth images, disparity images, and ground truth in our NPO dataset. Depth and disparity values increase from red to blue. ■ and ■ represent negative obstacles and positive obstacles, respectively.

an element-wise addition. Secondly, the output of the residual structure is fed into a global average pooling layer to adjust the resolution to 1×1. Thirdly, a Fully Connected (FC)-BN-ReLU layer and an FC layer are used to generate the weights of different channels of the fusion result. The FC-BN-ReLU layer reduces the number of channels by half, and the FC layer restores the number of channels. Fourthly, a Sigmoid layer is used to map the generated channel weights into [0,1]. The output of the residual structure is fused with the mapped weights through element-wise multiplication. Finally, a transposed Convolution-BN-ReLU layer is used to double the resolution. The output of the transposed Convolution-BN-ReLU layer is the output of each stage of the decoders.

D. The Loss Functions

To extract residual features of RGB feature maps from depth feature maps for semantic segmentation, the depth stream needs to have the ability to achieve semantic segmentation independently. So, in the training process, the loss between the ground truth y and the output of the depth stream $\hat{y}_d$ also needs to be calculated. We calculate the cross-entropy loss $\mathcal{L}_{seg}(y, \hat{y}_d)$ between the ground truth y and the output of the depth stream $\hat{y}_d$, as well as the cross-entropy loss $\mathcal{L}_{seg}(y, \hat{y}_{rgb})$ between the ground truth y and the output of the RGB stream $\hat{y}_{rgb}$, to train the InconSeg.

In the n-th RGF module, we use the cross-entropy loss $\mathcal{L}_{seg}(y_n, \hat{y}_n)$ between the ground truth y_n and RGB predicted mask $\hat{y}_n$ to guide the generation of RGB residual features y^n_{res}. We also use the cross-entropy loss $\mathcal{L}_{seg}(y^n_{res}, \hat{y}^n_{res})$ between the RGB residual features y^n_{res} and predicted residual features $\hat{y}^n_{res}$ to guide the extraction of RGB residual features from depth feature maps. So, the loss $\mathcal{L}^n_{RGF}$ of the n-th RGF module is represented as: $\mathcal{L}^n_{RGF} = \mathcal{L}_{seg}(y_n, \hat{y}_n) + \mathcal{L}_{seg}(y^n_{res}, \hat{y}^n_{res})$. The losses of each RGF module are also used to train our InconSeg. To sum up, the total loss $\mathcal{L}_{total}$ calculated as: $\mathcal{L}_{total} = \mathcal{L}_{seg}(y, \hat{y}_d) + \mathcal{L}_{seg}(y, \hat{y}_{rgb}) + \sum_{n=1}^{3} \mathcal{L}^n_{RGF}$. We use the $\mathcal{L}_{total}$ to train our InconSeg.

IV. The Released Dataset

A. Data Collection and Processing

As aforementioned, the multi-modal datasets with negative obstacles are very limited in the research community. Existing negative road-obstacle datasets are mainly small-scale and collected from urban scenes. So, in this work, we build and release a large-scale dataset with both negative and positive obstacles for road-obstacle segmentation.

The dataset is recorded with an on-vehicle ZED stereo camera in both urban and rural environments in Qingyuan of Fushun City, Liaoning Province, China. The camera and vehicle are shown in Fig. 4. We collect data on 7 different roads at different times to increase the diversity of the dataset. Fig. 4 shows several sample roadmaps for data collection. We totally collect 20 image sequences with 90,204 groups of images ($1,242 \times 2,208$ resolution). Each group contains a left RGB image, a right RGB image, and a 16-bit depth image produced by the ZED camera. The dataset contains various lighting conditions, such as normal lights, large-area shadows, dim lights, and sun glare. There are also different weather conditions, such as sunny, cloudy, and snowy. We use the LEAStereo [25] algorithm with left and right images to generate disparity images. We manually label *negative obstacles* (i.e., potholes and cracks) and *positive obstacles* (i.e., pedestrians, cars, and motorcycles) in 5,000 images.

We name our dataset Negative and Positive Obstacle (NPO) dataset. To the best of our knowledge, our NPO dataset is the largest dataset for semantic segmentation of road obstacles that include negative obstacles. Some sample images of our dataset are shown in Fig. 5.

B. Dataset Analysis

In our NPO dataset, there are 2,960 images collected from urban scenes and 2,040 images collected from rural scenes. The dataset can be divided into two types based on the conditions of the road surface: normal roads with 3,465 images and abnormal roads (e.g. snow, puddles) with 1,535 images. The images in this dataset are collected under a variety of weather conditions, that is, 4,231 images from sunny days, 583 images from snowy days,

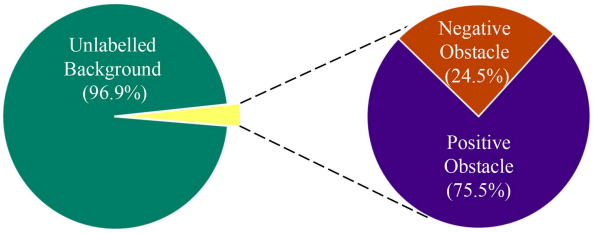


Fig. 6. Pixel ratio for each class in our dataset.

and 204 images from cloudy days. In our NPO dataset, 4,596 images include negative obstacles, and 3,105 images include positive obstacles. Fig. 6 shows the pixel ratio for each class.

V. Experimental Results and Discussions

A. The Datasets

1) NPO Dataset: We randomly split NPO dataset into three sets: training (2,500 groups of RGB-D images), validation (1,250 groups of RGB-D images), and testing (1,250 groups of RGB-D images). The resolution of the images is reduced to 288×512 during training and testing.

2) Pothole-600 Dataset: To verify the ability of our network to tackle inconsistency in multi-modal data, we randomly add interference information to disparity images in Pothole-600 dataset [6], so that inconsistent information exists in the RGB images and disparity images. We use the Pothole-600 dataset with random inconsistent information to validate the performance of our network. We also resize the images to the same resolution as the images in the NPO dataset.

B. Training Details

We implement our InconSeg with PyTorch. The network is trained and tested on a PC with NVIDIA RTX 3090 graphics card. We use the pre-trained weight of ResNet [24] from PyTorch to initialize the parameters of the first four encoder stages. During training, we use the stochastic gradient descent optimizer. We set the initial learning rate, the momentum, and the decay strategy as 0.01, 0.90, and 0.95, respectively.

C. Ablation Study

1) Ablation on the Position of the RGF Module: We conduct experiments to select the best structure for our InconSeg. We place the RGF module at different positions in the RGB decoder to design several variants. Firstly, we use the simple element-wise addition to replace the RGF modules in InconSeg. Secondly, we design five variants, each one with a RGF module at different stages of the RGB decoder, for example, the RGF module is placed after the first stage of the RGB decoder in a variant. Finally, we sequentially place 2, 3, 4, and 5 RGF modules into the RGB decoder from the first stage, respectively. We use the metrics, mean Accuracy (mAcc), mean F-score (mF1), and mean Intersection-over-Union (mIoU) [3], over both negative and positive obstacles to quantitatively evaluate the performance of the variants. The details for all variants and their results are displayed in Table I. We can see that the variant without any RGF module presents the worst result. This demonstrates that our proposed RGF module is beneficial to InconSeg. Comparing variants B with G, we can find that the performance improvement

TABLE I
Results (%) of the Ablation Study on the Position of Our RGF Module

No.	Fusion module					RGB+Disparity			RGB+Depth		
	5-th	4-th	3rd	2nd	1st	mAcc	mIoU	mF1	mAcc	mIoU	mF1
(A)	–	–	–	–	–	87.73	80.25	88.54	86.27	80.91	88.99
(B)	–	–	–	–	✓	88.83	83.06	90.40	88.56	83.00	90.37
(C)	–	–	–	✓	–	87.76	82.53	90.05	88.16	82.74	90.21
(D)	–	–	✓	–	–	87.94	82.40	89.98	88.37	82.56	90.07
(F)	–	✓	–	–	–	88.15	82.00	89.73	88.21	82.42	90.02
(G)	✓	–	–	–	–	87.67	81.82	89.63	88.34	82.38	89.95
(H)	–	–	–	✓	✓	89.21	82.58	90.15	87.93	82.94	90.35
(I)	–	–	✓	✓	✓	**89.44**	**83.88**	**90.96**	**89.65**	**83.76**	**90.89**
(J)	–	✓	✓	✓	✓	87.82	82.02	89.79	86.55	81.39	89.37
(K)	✓	✓	✓	✓	✓	88.59	81.55	89.51	85.99	81.11	89.24

The bold font represents the best result.

TABLE II
Results (%) of the Ablation Study on the Structure of the RGF Module

Method	RGB+Disparity			RGB+Depth		
	mAcc	mIoU	mF1	mAcc	mIoU	mF1
$\lvert R-D\rvert + \lvert D-R\rvert$	88.34	83.02	90.42	89.03	83.43	90.65
$\lvert R-D\rvert$	83.53	80.35	88.60	83.06	79.82	88.21
$\lvert D-R\rvert$	86.56	81.83	89.67	88.59	83.33	90.62
$R+D$	86.93	82.29	89.93	88.66	82.65	90.17
$R \odot D$	83.89	80.87	88.92	86.68	82.70	90.16
$R-D$	88.08	83.15	90.48	86.34	82.72	90.17
$D-R$	**89.44**	**83.88**	**90.96**	**89.65**	**83.76**	**90.89**

The bold font represents the best result.

brought by our RGF module gradually decreases as the resolution of the input data of RGF module decreases. The reason may be that the smaller the resolution of the feature maps, the less information the feature maps contain. So, the RGF module extracts fewer residual features from small-resolution feature maps. Comparing variants H, J, I, and K, we find that variant I with three RGF modules presents the best results in terms of all metrics. According to the experimental results, we choose variant I as our InconSeg (see Fig. 2).

2) Ablation on the Structure of RGF Module: For the structure of the RGF module, we design several variants with different methods to generate different features between RGB feature maps and depth feature maps. We use the following 7 methods to generate different features: 1) Subtract depth feature maps from RGB feature maps, denoted as $R - D$; 2) Subtract RGB feature maps from depth feature maps, denoted as $D - R$; 3) Add RGB feature maps and depth feature maps, denoted as $R + D$; 4) Concatenate RGB feature maps with depth feature maps, denoted as $R \odot D$; 5) Calculate the absolute value of the result of the first method, denoted as $|R - D|$; 6) Calculate the absolute value of the result of the second method, denoted as $|D - R|$; 7) Add the 5-th and 6-th methods, denoted as $|R - D| + |D - R|$. The purpose of the 5-th and 6-th is to turn negative values into positive ones in the different features generated by the first and second methods. The details and results of variants are displayed in Table II. We can find that the $D - R$ method presents the best performance. We also find that the negative values of different features are beneficial to the generation of residual features. According to the experimental results, we use the $D - R$ method to generate different features in the RGF module.

3) Ablation on Modality: We design several variants to demonstrate that our InconSeg performs better than using a single modality when fusing inconsistent RGB-D images. we design single-modal variants by removing the RGB stream or the depth stream from InconSeg. The single-modal variants are trained by RGB, depth, and disparity images, respectively. We

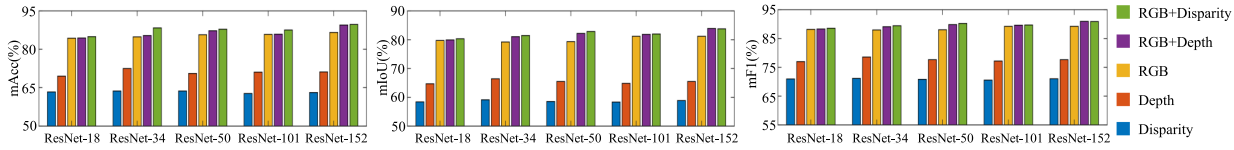


Fig. 7. Results of the ablation on modality. We can see that our InconSeg achieves better results with multi-modal fusion than with one-single modality.

TABLE III
COMPARATIVE RESULTS (%) ON THE TESTING SET OF OUR NPO DATASET

Method	Modality	Negative obstacle			Positive obstacle			mAcc	mIoU	mF1	RTX 3060	RTX 3090
		Acc	IoU	F1	Acc	IoU	F1				ms	ms
DeepLabV3+ [26]	RGB+Disparity	77.40	69.96	82.33	95.17	91.92	95.79	86.28	80.94	89.06	20.18	17.99
	RGB+Depth	76.00	69.43	81.96	94.77	91.60	95.62	85.39	80.52	88.79	20.21	18.01
	RGB	76.82	70.43	82.65	94.93	92.07	95.87	85.88	81.25	89.26	20.02	15.73
	Disparity	41.03	36.37	53.34	88.11	82.54	90.43	64.57	59.45	71.88	19.99	15.72
	Depth	57.25	50.96	67.51	87.71	82.93	90.67	72.48	66.94	79.09	19.97	15.80
FuseNet [7]	RGB+Disparity	60.24	56.09	71.87	92.54	87.70	93.45	76.39	71.90	82.66	43.56	17.62
	RGB+Depth	62.91	58.85	74.09	86.30	85.01	91.90	74.60	71.93	82.99	43.46	17.80
	RGB	74.19	68.04	80.98	95.11	92.07	95.87	84.65	80.06	88.43	29.70	11.58
	Disparity	9.66	9.59	17.51	80.69	65.79	79.37	45.18	37.69	48.44	29.63	11.57
	Depth	29.16	28.28	44.10	56.89	56.03	71.82	43.02	42.16	57.96	29.61	11.55
AA-RTFNet [6]	RGB+Disparity	76.61	67.96	80.92	94.47	91.90	95.78	85.54	79.93	88.35	77.38	50.76
	RGB+Depth	75.13	66.19	79.66	93.60	90.42	94.97	84.36	78.31	87.31	77.39	50.72
	RGB	76.80	67.93	80.90	94.91	92.09	95.88	85.86	80.01	88.39	54.21	31.88
	Disparity	41.02	36.13	53.08	91.24	85.06	91.93	66.13	60.60	72.51	54.29	32.32
	Depth	55.36	47.09	64.03	89.76	84.41	91.55	72.56	65.75	77.79	54.65	32.00
ESANet [8]	RGB+Disparity	76.34	68.99	81.65	94.61	91.38	95.49	85.47	80.18	88.57	25.19	21.80
	RGB+Depth	76.72	68.82	81.53	94.21	91.47	95.54	85.46	80.14	88.54	25.15	21.73
	RGB	78.90	68.42	81.25	93.56	90.59	95.06	86.23	79.51	88.16	16.36	14.77
	Disparity	40.53	34.13	50.89	86.97	81.51	89.81	63.75	57.82	70.35	16.38	14.71
	Depth	55.22	46.01	63.03	88.26	81.00	89.50	71.74	63.51	76.26	16.30	14.70
MAFNet [3]	RGB+Disparity	76.02	67.09	80.30	95.19	91.86	95.76	85.60	79.48	88.03	92.58	63.07
	RGB+Depth	75.97	65.57	79.21	93.17	90.38	94.95	84.57	77.98	87.08	92.33	63.15
	RGB	78.11	66.84	80.13	94.89	91.22	95.41	86.50	79.03	87.77	59.28	36.84
	Disparity	25.10	22.94	37.32	87.48	82.16	90.20	56.29	52.55	63.76	58.99	36.13
	Depth	14.21	13.56	23.88	89.70	84.24	91.45	51.95	48.90	57.66	58.73	36.24
RFNet [14]	RGB+Disparity	72.38	65.22	78.95	93.44	90.35	94.93	82.91	77.79	86.94	9.40	8.59
	RGB+Depth	73.02	65.79	79.37	91.35	88.22	93.74	82.19	77.01	86.55	9.75	8.60
	RGB	77.44	68.10	81.02	94.75	90.91	95.24	86.10	79.51	88.13	5.29	5.02
	Disparity	40.74	36.05	52.99	86.19	80.59	89.25	63.47	58.32	71.12	**5.28**	**4.98**
	Depth	55.64	49.33	66.07	86.42	80.51	89.20	71.03	64.92	77.63	5.25	4.99
CMoDE [17]	RGB+Disparity	59.29	55.05	71.01	89.77	87.38	93.27	74.53	71.22	82.14	43.54	25.99
	RGB+Depth	60.14	55.57	71.44	89.90	87.65	93.42	75.02	71.61	82.43	43.48	26.02
	RGB	57.65	53.18	69.44	90.96	87.59	93.39	74.31	70.39	81.41	20.84	14.95
	Disparity	31.27	29.36	45.39	79.00	75.72	86.18	55.14	52.54	65.79	20.81	14.86
	Depth	44.62	41.05	58.21	79.29	75.60	86.11	61.95	58.33	72.16	20.78	14.94
SA-Gate [16]	RGB+Disparity	67.39	59.69	74.75	91.38	87.90	93.56	79.39	73.79	84.16	52.32	36.40
	RGB+Depth	56.93	50.59	67.19	89.69	85.78	92.35	73.31	68.19	79.77	52.43	36.39
TokenFusion [20]	RGB+Disparity	82.44	72.04	83.75	**96.57**	92.79	96.26	89.51	82.42	90.01	124.95	109.39
	RGB+Depth	**83.56**	73.48	84.71	96.50	92.86	93.60	**90.03**	83.17	89.16	120.01	105.10
InconSeg (Ours)	RGB+Disparity	82.76	**74.63**	**85.47**	96.12	**93.14**	**96.45**	89.44	**83.88**	**90.96**	76.39	47.15
	RGB+Depth	83.43	74.51	85.39	95.87	93.01	96.38	89.65	83.76	90.89	76.50	47.63

The bold font represents the best result.

compare the results of these variants with those of the multi-modal fusion InconSeg. We also use ResNet-18, ResNet-34, ResNet-50, ResNet-101, ResNet-152 as the backbone in the encoders to design several variants. The results of these variants are shown in Fig. 7.

The variant with the RGB modality achieves the best results among the single-modal variants. In the variants with the same encoder, we can find that the variants fusing two modalities perform better than those with only one modality. This shows that the fusion strategy of InconSeg is effective with different encoders. The results demonstrate that InconSeg can achieve better results than a single modality even when fusing two modalities of data with inconsistent information.

D. Comparative Study

We compare our proposed InconSeg with well-known networks: DeepLabV3+ [11], FuseNet [7], AA-RTFNet [6], MAFNet [3], ESANet [8], RFNet [14], CMoDE [17], SA-Gate [16], TokenFusion [20]. The first seven networks are trained with three single-modal data (i.e., RGB, depth, disparity) and two multi-modal data (i.e., RGB-depth data and RGB-disparity data), respectively. The multi-modal networks are trained with single-modal images by removing the RGB stream or the depth stream, except SA-Gate and TokenFusion. The compared networks can be divided into 3 types according to fusing strategy: 1) Fusing with sample concatenation and addition, including DeepLabV3+, FuseNet, and AA-RTFNet; 2) Fusing with attention module, including ESANet, MAFNet, and RFNet; 3) Fusing with specialized modules, including CMoDE, SA-Gate, and TokenFusion. We use the metrics, Acc, F1, and IoU for negative and positive obstacles, as well as mAcc, mF1, and mIoU, to quantitatively evaluate the overall performance of the networks. We also test the inference speed for each network on RTX 3060 and RTX 3090.

1) The Overall Results on Our NPO Dataset: Table III displays the results of all networks trained and tested on our NPO dataset. Comparing the results of DeepLabV3+, FuseNet, and AA-RTFNet, we can find that the best performance is achieved when the input is the RGB modality data. In

TABLE IV
COMPARATIVE RESULTS (%) ON DIFFERENT SCENES IN THE TESTING SET OF OUR NPO DATASET

Method	Modality	Abnormal road			Normal road			Urban scenes			Rural scenes		
		mAcc	mIoU	mF1	mAcc	mIoU	mF1	mAcc	mIoU	mF1	mAcc	mIoU	mF1
AA-RTFNet [6]	RGB+Disparity	79.05	73.13	84.41	85.57	79.83	88.26	84.94	79.14	87.78	84.04	77.64	87.23
FuseNet [7]	RGB+Disparity	63.72	59.59	74.68	75.97	71.21	82.01	75.98	71.03	81.93	69.50	66.04	79.33
MAFNet [3]	RGB+Disparity	81.15	74.81	85.48	85.32	78.86	87.56	85.06	78.49	87.32	83.10	75.94	86.21
ESANet [8]	RGB+Disparity	80.02	73.62	84.76	85.42	79.91	88.35	85.30	79.76	88.27	80.57	75.23	85.78
RFNet [14]	RGB+Disparity	73.23	68.08	80.99	83.00	77.77	86.89	82.98	77.33	86.60	78.75	73.11	84.31
SA-Gate [16]	RGB+Disparity	67.56	61.36	76.05	79.30	73.87	84.16	78.90	73.18	83.66	72.15	66.90	80.05
CDoME [17]	RGB+Disparity	65.77	61.78	76.24	74.31	70.95	81.85	73.39	69.94	81.05	75.39	71.45	82.73
TokenFusion [20]	RGB+Disparity	**85.38**	76.55	86.69	89.17	81.91	89.64	88.97	81.73	88.36	**89.09**	79.34	88.39
InconSeg (Ours)	RGB+Disparity	82.28	**77.13**	**87.08**	**89.79**	**83.84**	**90.91**	**89.27**	**83.37**	**90.62**	87.92	**79.91**	**88.79**
AA-RTFNet [6]	RGB+Depth	76.37	69.31	81.86	83.91	78.00	87.05	83.49	77.38	86.63	82.82	76.47	86.47
FuseNet [7]	RGB+Depth	56.07	54.14	70.06	75.01	71.95	82.92	74.86	71.62	82.74	67.46	65.52	79.03
MAFNet [3]	RGB+Depth	75.44	68.97	81.63	84.37	77.52	86.70	83.72	76.88	86.26	82.34	75.93	86.14
ESANet [8]	RGB+Depth	81.10	75.53	85.98	84.96	79.53	88.08	84.78	79.47	88.06	80.93	74.80	85.53
RFNet [14]	RGB+Depth	71.90	66.72	80.04	82.10	76.86	86.40	81.87	76.31	86.04	79.33	72.16	83.76
SA-Gate [16]	RGB+Depth	59.25	54.00	70.13	72.44	67.40	78.95	72.01	66.58	78.30	68.17	62.55	76.75
CDoME [17]	RGB+Depth	64.63	59.99	74.96	75.85	72.26	82.78	74.78	71.13	81.91	75.49	71.07	82.60
TokenFusion [20]	RGB+Depth	**86.55**	**78.72**	**88.07**	**89.76**	82.52	90.05	89.30	82.10	89.79	**89.95**	80.78	89.31
InconSeg (Ours)	RGB+Depth	82.59	76.92	86.94	89.66	**83.62**	**90.77**	**89.44**	**83.37**	**90.63**	88.49	**81.42**	**89.65**

The bold font represents the best result.

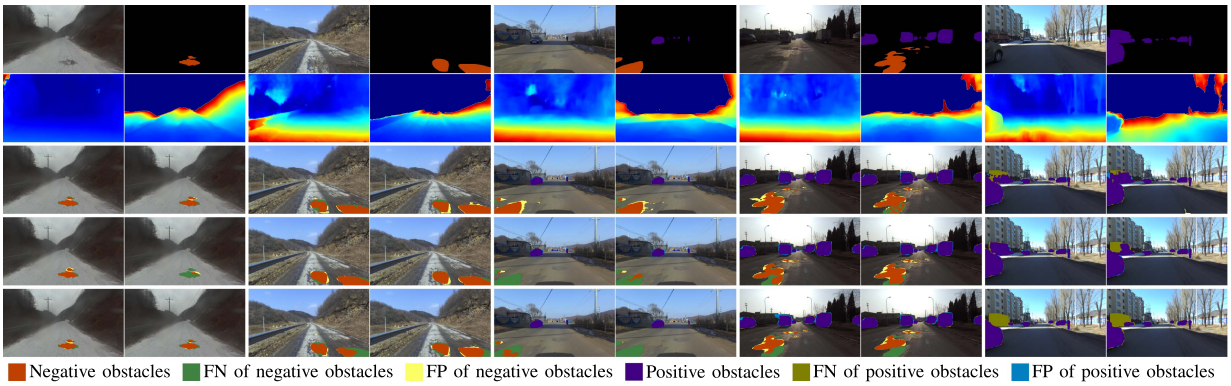


Fig. 8. Sample qualitative results for the multi-modal networks. The odd and even columns in the first row represent the RGB image and their ground truth, respectively. The odd and even columns in the second row represent disparity and depth images, respectively. The odd and even columns of the remaining rows represent the results of different networks with RGB-disparity and RGB-depth, respectively. The third row to the last row are respectively the results of our InconSeg, TokenFusion [20], and ESANet [8]. FN and FP represent *false negative* and *false positive*, respectively.

contrast, the performance degrades when the input is the inconsistent RGB-D modalities of data. This demonstrates that fusing two modalities with inconsistent information using a simple channel-concatenation or element-wise addition method is not the optimal fusion method. The other compared networks adopt attention modules or specialized-design modules to fuse multi-modal features. However, the results of these networks with multi-modal data are only slightly higher than the RGB modality (e.g., ESANet, CMoDE), and some results are even worse than the RGB modality. According to section V.C.3, our InconSeg achieves better results than using a single modality (higher 1.88% and 2%) when fusing multi-modal data with inconsistent information. Comparing all the results, our InconSeg achieves the best performance, both in terms of fusing RGB-depth images or fusing RGB-disparity images, which demonstrates the superiority of our InconSeg. The results of the inference speed show that our InconSeg achieves an acceptable inference speed.

2) The Results in Different Scenes: We also evaluate the multi-modal networks using images from different scenes. Table IV displays the results of all the multi-modal networks. According to the road conditions, we divide the testing set into two subsets: normal roads, and abnormal roads with snow or water. Similarly, according to the environment, we divide the testing set into two subsets: urban scenes and rural scenes. Comparing all the network results of abnormal roads and normal roads, we can find that the water and snow on the road have a large impact on results. Similarly, the results for urban scenes and rural scenes show that obstacle segmentation in rural scenes is more challenging than in urban scenes. We can see that our InconSeg achieves the best performance in all the scenarios, except the abnormal-road scene, in terms of whether fusing RGB and depth images or fusing RGB and disparity images.

3) The Qualitative Results: Some sample qualitative results for the top-3 multi-modal networks in Table III are shown in Fig. 8. From the 5-th and 6-th columns, we can find that it is a challenge to segment the negative obstacles when the negative obstacles have a similar texture to the road. From the last four columns, we can find that shadows with large areas and oncoming sunshine are challenging for obstacle segmentation. However, Fig. 8 illustrates that our proposed InconSeg achieves state-of-the-art performance in these challenging scenarios. Fig. 8 also illustrates that the networks trained with RGB images and disparity images have similar results as the networks trained with RGB images and depth images.

4) The Overall Results on the Pothole-600 Dataset: The results of all networks on the Pothole-600 dataset are displayed in Table V. From the results, we can find that the multi-modal fusion results of some networks are better than those using a single modality. This indicates that these networks also have some ability to overcome the negative effects of inconsistency

TABLE V
COMPARATIVE RESULTS (%) ON THE AUGMENTED POTHOLE-600 DATASET

Method	Modality	Potholes			3060	3090
		Acc	IoU	F1	ms	ms
DeepLabV3+ [26]	RGB+Disparity	76.04	54.76	70.77	20.42	18.28
	RGB	66.11	54.77	70.78	19.97	15.63
	Disparity	67.80	54.50	70.55	19.85	15.87
FuseNet [7]	RGB+Disparity	63.81	41.50	58.66	43.87	17.88
	RGB	61.30	47.78	64.66	29.63	11.45
	Disparity	72.64	42.81	59.95	29.64	11.68
AA-RTFNet [6]	RGB+Disparity	74.00	54.80	70.80	77.55	50.88
	RGB	71.39	55.05	71.01	54.39	32.23
	Disparity	39.45	34.31	51.09	54.09	31.91
ESANet [8]	RGB+Disparity	74.57	56.20	71.96	25.25	22.13
	RGB	**77.74**	53.34	69.57	16.85	14.45
	Disparity	56.29	47.19	64.12	16.67	14.94
MAFNet [3]	RGB+Disparity	18.42	17.56	29.87	92.58	63.45
	RGB	34.43	30.32	46.54	59.12	36.74
	Disparity	8.37	8.21	15.17	59.01	36.56
RFNet [14]	RGB+Disparity	70.34	51.67	68.13	9.96	8.97
	RGB	64.25	52.37	68.74	**5.27**	**5.00**
	Disparity	67.62	52.14	68.54	5.49	5.03
CMoDE [17]	RGB+Disparity	51.83	44.47	61.56	43.51	25.89
	RGB	53.99	41.45	58.60	20.75	14.92
	Disparity	47.32	42.53	59.68	20.86	14.91
SA-Gate [16]	RGB+Disparity	46.79	39.49	56.62	52.77	36.81
TokenFusion [20]	RGB+Disparity	73.79	59.54	74.64	129.21	110.61
InconSeg (Ours)	RGB+Disparity	75.94	**61.24**	**75.96**	76.67	47.46
	RGB	73.28	54.03	70.16	36.70	24.34
	Disparity	61.44	52.10	68.51	36.63	24.29

The bold font represents the best result.

in multi-modal data. However, the multi-modal fusion results of our InconSeg are significantly better than the single-modal results. This indicates that our proposed InconSeg can better solve the problem of inconsistency in multi-modal data. Comparing all the results, our InconSeg achieves the best results, which also shows the superiority of our InconSeg.

VI. CONCLUSION AND FUTURE WORK

We proposed here a novel network with RGF modules for the segmentation of negative and positive road obstacles. Our proposed network addressed the performance degradation when fusing two modalities with inconsistent information. We utilized two independent data streams to extract features and predict masks from RGB modality and depth modality, respectively. The RGF module is used to extract and fuse the residual features of RGB images from the output of the stages of the depth decoder. In addition, we released a large-scale RGB-D dataset with pixel-level labels of negative and positive road obstacles to verify the performance of our network. The experimental results demonstrate the superiority of our network. The results also demonstrate that our network can achieve better results than a single modality when fusing multi-modal data with inconsistent information. However, our proposed network still has several limitations, for example, segmentation results on the edge of negative obstacles are inaccurate. So, in the future, we would like to introduce the edge information of obstacles to improve the segmentation performance.

REFERENCES

[1] S. Wang, Y. Sun, Z. Wang, and M. Liu, "St-TrackNet: A multiple-object tracking network using spatio-temporal information," *IEEE Trans. Automat. Sci. Eng.*, early access, 2022, doi: 10.1109/TASE.2022.3216450.

[2] P. Cai, H. Wang, Y. Sun, and M. Liu, "DQ-GAT: Towards safe and efficient autonomous driving with deep Q-learning and graph attention networks," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 11, pp. 21102–21112, Nov. 2022.

[3] Z. Feng et al., "MAFNet: Segmentation of road potholes with multimodal attention fusion network for autonomous vehicles," *IEEE Trans. Instrum. Meas.*, vol. 71, 2022, Art. no. 3523712.

[4] K. Muhammad et al., "Vision-based semantic segmentation in scene understanding for autonomous driving: Recent achievements, challenges, and outlooks," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 12, pp. 22694–22715, Dec. 2022.

[5] N. Ma et al., "Computer vision for road imaging and pothole detection: A state-of-the-art review of systems and algorithms," *Transp. Saf. Environ.*, vol. 4, no. 4, Nov., 2022, Art. no. tdac026.

[6] R. Fan, H. Wang, M. J. Bocus, and M. Liu, "We learn better road pothole detection: From attention aggregation to adversarial domain adaptation," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 285–300.

[7] C. Hazirbas, L. Ma, C. Domokos, and D. Cremers, "FuseNet: Incorporating depth into semantic segmentation via fusion-based CNN architecture," in *Proc. Asian Conf. Comput. Vis.*, 2016, pp. 213–228.

[8] D. Seichter, M. Köhler, B. Lewandowski, T. Wengefeld, and H.-M. Gross, "Efficient RGB-D semantic segmentation for indoor scene analysis," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2021, pp. 13525–13531.

[9] D. Feng et al., "Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 3, pp. 1341–1360, Mar. 2021.

[10] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*.

[11] 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.*, 2018, pp. 801–818.

[12] R. Azad et al., "Transdeeplab: Convolution-free transformer-based deeplab v3+ for medical image segmentation," in *Proc. Int. Workshop PRedictive Intell. I Med.*, 2022, pp. 91–102.

[13] Y. Sun, W. Zuo, and M. Liu, "RTFNet: RGB-thermal fusion network for semantic segmentation of urban scenes," *IEEE Robot. Autom. Lett.*, vol. 4, no. 3, pp. 2576–2583, Jul. 2019.

[14] 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. Automat. Lett.*, vol. 5, no. 4, pp. 5558–5565, Oct. 2020.

[15] X. Ying and M. C. Chuah, "UCTNet: Uncertainty-aware cross-modal transformer network for indoor RGB-D semantic segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 20–37.

[16] X. Chen et al., "Bi-directional cross-modality feature propagation with separation-and-aggregation gate for RGB-D semantic segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 561–577.

[17] A. Valada, J. Vertens, A. Dhall, and W. Burgard, "AdapNet: Adaptive semantic segmentation in adverse environmental conditions," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2017, pp. 4644–4651.

[18] A. Valada, G. L. Oliveira, T. Brox, and W. Burgard, "Deep multispectral semantic scene understanding of forested environments using multimodal fusion," in *Proc. Int. Symp. Exp. Robot.*, 2017, pp. 465–477.

[19] A. Pfeuffer and K. Dietmayer, "Robust semantic segmentation in adverse weather conditions by means of fast video-sequence segmentation," in *Proc. IEEE 23rd Int. Conf. Intell. Transp. Syst.*, 2020, pp. 1–6.

[20] Y. Wang, X. Chen, L. Cao, W. Huang, F. Sun, and Y. Wang, "Multimodal token fusion for vision transformers," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 12186–12195.

[21] 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. Int. Conf. Intell. Robots Syst.*, 2016, pp. 1099–1106.

[22] X. Han, C. Nguyen, S. You, and J. Lu, "Single image water hazard detection using FCN with reflection attention units," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 105–120.

[23] M. Bijelic et al., "Seeing through fog without seeing fog: Deep multimodal sensor fusion in unseen adverse weather," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 11682–11692.

[24] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[25] X. Cheng et al., "Hierarchical neural architecture search for deep stereo matching," in *Proc. 34th Int. Conf. Neural Inf. Proc. Syst.*, vol. 33, pp. 22158–22169, 2020.