Segmentation of Road Negative Obstacles Based on Dual Semantic-Feature Complementary Fusion for Autonomous Driving

Zhen Feng, *Member, IEEE*, Yanning Guo, and Yuxiang Sun, *Member, IEEE*

***Abstract*—Segmentation of road negative obstacles (i.e., potholes and cracks) is important to the safety of autonomous driving. Although existing RGB-D fusion networks could achieve acceptable performance, most of them only conduct binary segmentation for negative obstacles, which does not distinguish potholes and cracks. Moreover, their performance is susceptible to depth noises, in which case the fluctuations of depth data caused by the noises may make the networks mistakenly treat the area as a negative obstacle. To provide a solution to the above issues, we design a novel RGB-D semantic segmentation network with dual semantic-feature complementary fusion for road negative obstacle segmentation. We also re-label an RGB-D dataset for this task, which distinguishes road potholes and cracks as two different classes. Experimental results show that our network achieves state-of-the-art performance compared to existing well-known networks.**

***Index Terms*—Semantic segmentation, RGB-D fusion, negative obstacles, potholes, cracks, autonomous driving.**

I. Introduction

In autonomous driving, semantic image segmentation is one of the fundamental tasks for environment perception [1], [2], [3], [4]. It aims to classify objects at the pixel level. One of the major applications of semantic segmentation is to detect obstacles on roads, such as positive obstacles [5], [6], [7], [8] and negative obstacles (i.e., potholes and cracks) [9], [10]. Road negative obstacles impose threats to the safety of autonomous driving. It could lead to abrupt vibrations, or even rollovers to vehicles if the vehicle speed is fast [11]. So, accurate segmentation of potholes and cracks is important for autonomous driving.

Existing methods have used single RGB images or single thermal images to segment negative obstacles [12], [13]. Since negative obstacles are below road surfaces, their height information would benefit the segmentation. So, some researchers resort to using depth data, such as point clouds [14], depth images, and disparity images [15]. To further improve segmentation performance, some methods fuse depth data with RGB data [16], [17]. However, most existing methods only conduct binary segmentation for negative obstacles. They do not distinguish potholes and cracks [18], [19], which would be important for downstream tasks. For example, potholes usually have more impact on vehicle safety compared to cracks, especially when the vehicle speed is fast, so vehicles should pay more attention to potholes. Moreover, most methods use stereo cameras to obtain depth data, which would become noisy for far distances or surfaces without textures. Negative obstacles are usually shallow, so the depth data fluctuations caused by noises would make existing networks mistakenly treat noisy areas as negative obstacles. Especially for perspective-view images, the depth data are more prone to be influenced by those from down-looking cameras. Fig. 1 shows an example for depth fluctuations caused by noises.

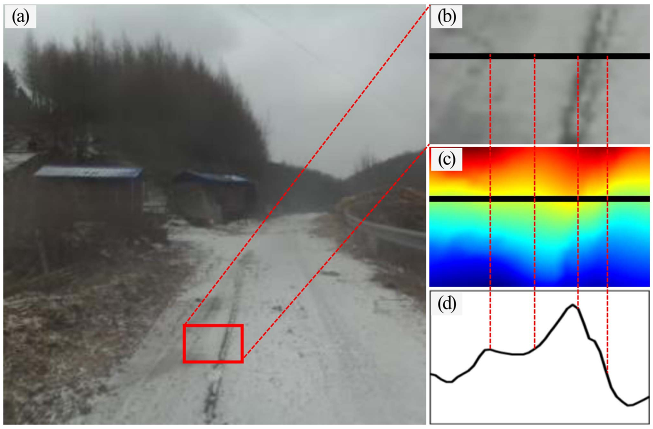

Fig. 1. Example of depth data fluctuations caused by noises. (b) Is the zoom-in view of the patch in (a). (c) Is the corresponding depth image of (b), which is visualized in the *jet* color map. (d) Is the plot of the depth values of the black line in (c).

To address the above issues, we propose a dual semantic-feature complementary fusion strategy, which does not directly fuse all features from the two modalities. Specifically, we place the fusion stage at the end of the decoder. This structure ensures that the noises in the depth image would not affect the

Manuscript received 16 January 2024; revised 16 February 2024; accepted 24 February 2024. Date of publication 18 March 2024; date of current version 12 June 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 62003286, in part by Guangdong Basic and Applied Basic Research Foundation under Grant 2022A1515010116, and in part by the City University of Hong Kong under Grant 9610675. *(Corresponding author: Yuxiang Sun.)*

Zhen Feng is with the Department of Mechanical Engineering, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong (e-mail: zhenfeng@polyu.edu.hk).

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

Yuxiang Sun is with the Department of Mechanical Engineering, City University of Hong Kong, Kowloon, Hong Kong (e-mail: yx.sun@cityu.edu.hk).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TIV.2024.3376534.

Digital Object Identifier 10.1109/TIV.2024.3376534

extraction of RGB features in the encoding stage. In addition, our semantic feature complementary fusion strategy extracts only the semantic features from the other model, which are missing when segmenting negative obstacles. The method of extracting missing semantic features can avoid the influences of the noises in the depth images, since the noises may have few semantic features of objects. Moreover, we re-label our previously released NPO dataset [5] with the two classes (i.e., potholes and cracks) separated. Models trained with the new labels would be able to discriminate the two classes. The main contributions of this work are summarized as follows:

1) We design a Dual Semantic-feature Complementary Fusion (DSCF) module to extract complementary semantic features for each modality.
2) We propose a novel RGB-D fusion network named PotCrackSeg with the DSCF module for the segmentation of potholes and cracks.
3) We upgrade the existing NPO dataset by re-labeling the new classes (i.e., potholes and cracks). Our code and dataset are open-sourced.[1]

This paper is structured as follows. Section II reviews the related work. Section III describes our proposed network. Section IV presents our re-labeled dataset. Section V discusses the experimental results. Conclusions and future work are drawn in the last section.

II. Related Work

A. Strategies for Multi-Modal Fusion

Directly concatenation and element-wise addition have been widely used to fuse multi-modal data [20], [21]. To improve the fusion strategy, many researchers have used attention modules for fusion. Zhou et al. [22] designed a cross-modal attention fusion module to fuse the complementary information from RGB and thermal images. The authors adopted two streams to exploit the complementary information with multi-head cross attention modules, and fused the outputs of the two streams by element-wise addition. Zhou et al. [23] combined modality attention, spatial attention, and correlation attention to design a tri-attention fusion module to fuse the features. Liang et al. [24] proposed EAEFNet with explicit attention-enhanced fusion modules to fuse the features from two modalities.

Many researchers have designed fusion modules with different fusion strategies. Seichter et al. [25] designed ESANet that re-weights features with the squeeze and excitation module before fusing them. Zhou et al. [26] proposed a deep feature fusion module with dense structure and convolutions with different dilation to fuse features. Zhou et al. [27] designed a cross-modal awareness module to fuse features with three stages. Wang et al. [28] proposed TokenFusion with a transformer structure to generate fusion tokens of different modalities. Zhou et al. [29] proposed a hierarchical multi-modal fusion module to fuse RGB and thermal images. Wang et al. [30] proposed a semantic-guided fusion module in SGFNet, using semantic information to guide the fusion process. Zhou et al. [31] proposed a memory sharing module to fuse features with three branches. Feng et al. [5] proposed a residual-guided fusion module to extract missing features of RGB images from depth data. Yang et al. [32] proposed a semantic-modulated cross-modal interaction mechanism to fuse different modalities

B. Datasets for Road Negative Obstacles

Liu et al. [33] released the DeepCrack dataset for the detection and segmentation of road cracks. There are 537 RGB images with the 544×384 resolution. Nguyen et al. [34] built the 2StagesCrack dataset that consists of 2,000 images with the 96×96 resolution. Rateke et al. [35] built a segmentation dataset on the RTK dataset [36], which contains RGB images of traffic scenes. They manually label 701 images in different scenes. They labeled 11 classes, including potholes and cracks. Guo et al. [37] built the UDTIRI dataset that contains 1,000 images, which are from online resources, existing datasets [38], and self-collected images. Han et al. [39] built the Puddle-1000 dataset with manually labeled puddles, including 985 RGB images captured by a ZED camera. Arya et al. [40] built a large-scale dataset RDD-2020 containing 26,620 images for the detection of negative obstacles including cracks and potholes. They used a smartphone to collect images from several countries under various weather and lighting conditions. Bhatia et al. [13] used a FLIR ONE thermal camera to acquire images under varying lighting conditions to build a road pothole segmentation dataset. They manually labeled 4,904 images augmented from captured 500 images with the resolution of 240×295. Fan et al. [16] released the Pothole-600 dataset that contains 600 pairs of RGB images and transformed disparity images with the resolution of 400×400 for the segmentation of road potholes. Feng et al. released the DRNO dataset [41] and the NPO dataset [5] for negative obstacles segmentation.

Although there are many datasets, datasets with multi-modal images are still scarce, and their scale is small. Moreover, the existing datasets only provide binary segmentation masks, without separating the two classes.

C. Segmentation Methods for Road Negative Obstacles

There are some works that focus on the segmentation of road cracks. Han et al. [42] designed CrackW-Net in a skip-level round-trip sampling block structure to address the misidentification caused by crack interruption and background noises. Chen et al. [43] designed LECSFormer with a proposed dense-structure decoder for crack detection. Sun et al. [10] proposed a multi-scale attention module and placed it in the decoder of DeeplabV3+ [44] for road crack segmentation. The authors also introduced dynamic weighting for high-level and low-level feature maps in the decoder. Fan et al. [45] designed Parallel ResNet for the segmentation of road cracks. Parallel ResNet consists of two proposed parallel ResNet modules. Each one contains three parallel residual convolutional branches and the outputs of the branches are fused by element-wise addition. Zhou et al. [46] proposed a lightweight network for crack detection by splitting the feature maps into high-resolution stage and low-resolution stage.

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

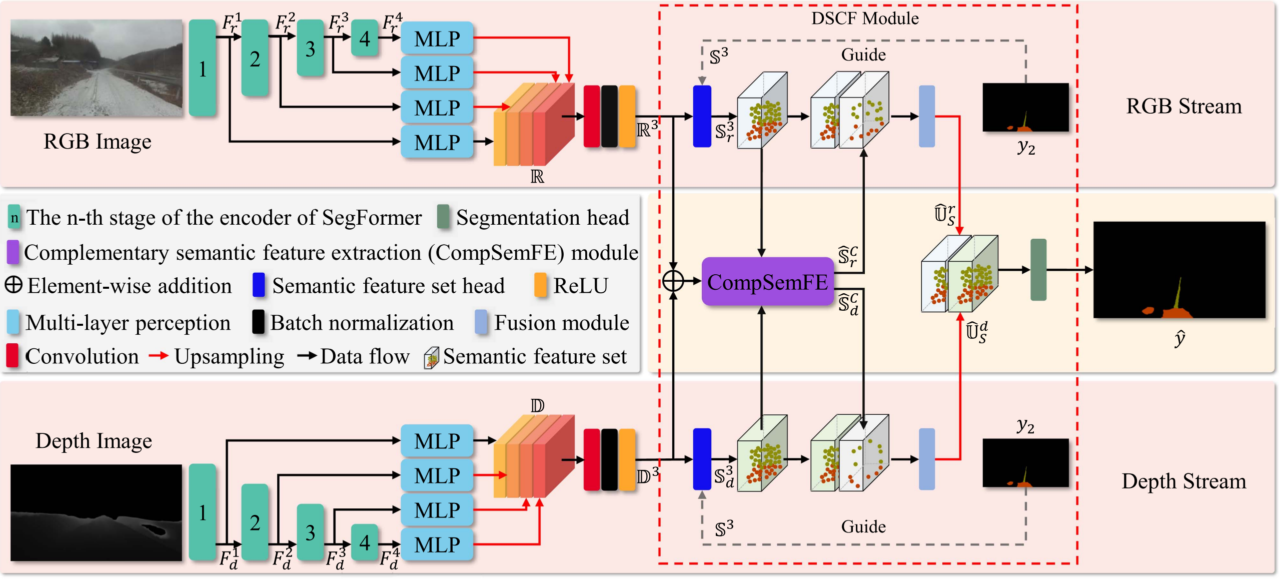

Fig. 2. Overall architecture of our proposed PotCrackSeg. It has an RGB stream and a depth stream. In each stream, multi-level features are extracted from the input image by a multi-stage encoder. The features are then fused and mapped into the semantic feature set. The complementary semantic features for each semantic feature set are extracted by the CompSemFE module and fed into each stream. The complemented semantic features from both streams are fed into a segmentation head to generate segmentation maps. The figure is best viewed in color.

There are also some works that focus on the segmentation of road potholes. For example, Masihullah et al. [9] designed a coupled framework for roads and potholes segmentation by combining DeeplabV3+ and an attention module. Fan et al. [16] designed AARTFNet and AAUNet by introducing attention modules between the encoder and decoder for road pothole segmentation. Feng et al. [47] designed MAFNet with two types of attention-based fusion modules for pothole segmentation by fusing RGB images and transformed disparity images. Katsamenis et al. [48] designed a network for road pothole segmentation based on the transformer structure. Liu et al. [49] introduced a coordinate attention module to UNet to design a network for the segmentation of both potholes and cracks. Feng et al. [41] proposed AMFNet for road and negative obstacles segmentation with an adaptive-mask fusion module. Feng et al. [5] proposed InconSeg to address inconsistencies between different modalities for the segmentation of positive obstacles and negative obstacles.

D. *Differences From Existing Methods*

Our network differs from the aforementioned networks in two aspects: 1) We segment both potholes and cracks and discriminate the two classes; 2) We indirectly fuse the two modalities of data. Specifically, we calculate the missing semantic features for each modality and extract that semantic features from the other modality for fusion.

III. The Proposed Network

A. *The Overall Architecture*

To alleviate the influence of the fluctuations caused by noises during the RGB-D fusion process, we propose a network named PotCrackSeg with a dual semantic-feature complementary fusion strategy. Specifically, we fuse the features with their missing semantic features instead of fusing all features containing noises since there are few semantic features of objects in the noises. So, our PotCrackSeg first extracts features from the two modalities and then calculates the missing semantic features for each modality and extracts the missing features. Finally, our PotCrackSeg fuses the features and their missing semantic features.

Fig. 2 shows the overall architecture of PotCrackSeg, which consists of an RGB stream, a depth stream, and a Dual Semantic-feature Complementary Fusion (DSCF) module. The DSCF module contains the last part of both streams, a Complementary Semantic Feature Extraction (CompSemFE) module, and a segmentation head. RGB images (denoted as I_r) and depth images (denoted as I_d) are fed into the RGB stream and the depth stream, respectively. The depth stream has the same architecture as the RGB stream. We adopt SegFormer [50] as the encoder for each stream to extract features from input data. There are four stages in the encoder of the RGB stream. The output of the n-th stage is denoted as F_r^n, where r refers to RGB and $n \in [1, 4]$. The resolution of F_r^n is $\frac{1}{2^{n+1}}$ of the resolution of the original input I_r. The outputs of each stage are processed by multi-layer perception (MLP) blocks, and then concatenated along the channel axis. Since the feature map resolutions are not the same, the outputs of the MLP blocks for the last three stages are resized by an upsampling layer, so that the feature maps have the same resolution before concatenation. The concatenation results contain all the features extracted by the RGB and depth encoders, which are called RGB and depth feature sets and denoted as $\mathbb{R}$ and $\mathbb{D}$:

$$\mathbb{R} = \{x | x = \mathrm{up}_r^n\left(\mathrm{mlp}_r^n(F_r^n)\right), n \in [1, 4]\}, \quad (1)$$

$$\mathbb{D} = \{x | x = \mathrm{up}_d^n \left(\mathrm{mlp}_d^n(F_d^n)\right), n \in [1, 4]\}, \tag{2}$$

where $\mathrm{mlp}_r^n(\cdot)$ refers to the n-th multi-layer perception block for the n-th stage output F_r^n, $\mathrm{up}_r^n(\cdot)$ refers to the n-th upsampling layer for the n-th MLP block. $\mathrm{up}_r^1(\cdot)$ refers to there being no operation on the input (i.e., no upsampling layer). Then, the channel number of $\mathbb{R}$ and $\mathbb{D}$ are reduced to 3 through a convolution layer, a batch normalization layer, and a ReLU layer. The number 3 means the number of classes to be segmented, including the unlabeled background. The outputs of the ReLU layers are denoted as $\mathbb{R}^3$ and $\mathbb{D}^3$. Both $\mathbb{R}^3$ and $\mathbb{D}^3$ are fed into the DSCF module. They are first mapped to the semantic feature set and fused with complementary semantic features of each modality. The output of the DSCF module $\hat{y}$ is the output of the whole PotCrackSeg.

B. The DSCF Module

The DSCF module consists of four parts: the last parts of both streams that are used to map features to the semantic feature set and complement semantic features; a CompSemFE module that is used to extract complementary semantic features; a segmentation head that is used to generate segmentation maps.

Firstly, the $\mathbb{R}^3$ is mapped to a semantic feature set through a semantic feature set head. In the semantic feature set, each element contains information about the class to which the pixel belongs. The semantic feature set is denoted as $\mathbb{S}^3$. The semantic feature set head consists of a convolution layer. To ensure that the semantic feature set head maps $\mathbb{R}^3$ to the semantic feature set where the negative obstacles are, semantic information about the negative obstacles is used to guide the mapping process. In our DSCF module, we adopt the ground truth as the guidance information. Since the resolution of $\mathbb{R}^3$ is $\frac{1}{4}$ of that of input image I_r, we downsample the ground truth with the nearest neighbor interpolation method to generate the guidance information y_2. The process is described as follows:

$$\mathbb{S}_r^3 \leftarrow \mathbb{R}^3 : Q(\mathbb{R}^3 | y_2) = Q(\mathrm{conv}_r(x) | y_2), \tag{3}$$

where $x \in \mathbb{R}^3$ and $\mathbb{S}_r^3$ is the RGB semantic feature set, which is in $\mathbb{S}^3$. $\mathrm{conv}_r(\cdot)$ refers to the process of the convolution layer, and $Q(i|k)$ refers to the guiding process that extracts the same semantic information from i using k as the supervisory information with a cross-entropy loss. The depth semantic feature set $\mathbb{S}_d^3$ is generated in the same way:

$$\mathbb{S}_d^3 \leftarrow \mathbb{D}^3 : Q(\mathbb{D}^3 | y_2) = Q(\mathrm{conv}_r(x) | y_2), \tag{4}$$

where $x \in \mathbb{D}^3$. After generating $\mathbb{S}_r^3$ and $\mathbb{S}_d^3$, both of them, $\mathbb{R}^3$, and $\mathbb{D}^3$ are fed into the CompSemFE module to extract the complementary sets ($\hat{\mathbb{S}}_r^C$ and $\hat{\mathbb{S}}_d^C$) in the set $\mathbb{S}^3$, where C means a complementary set containing complementary semantic features. $\mathbb{S}^3$ is generated from y_2. The RGB (depth) semantic feature set $\mathbb{S}_r^3$ ($\mathbb{S}_d^3$) and its complementary set $\mathbb{S}_r^C$ ($\mathbb{S}_d^C$) are concatenated along the channel axis and fused through a fusion module to generate fusion results $\hat{\mathbb{U}}_S^r$ ($\hat{\mathbb{U}}_S^d$). $\hat{\mathbb{U}}_S^r$ and $\hat{\mathbb{U}}_S^d$ are the outputs of the RGB stream and depth stream, respectively. Next, $\hat{\mathbb{U}}_S^r$ and $\hat{\mathbb{U}}_S^d$ are fused through concatenation along the channel axis after upsampling. The upsample layers resize the resolutions of $\hat{\mathbb{U}}_S^r$ and $\hat{\mathbb{U}}_S^d$ to that of original input images. Finally, the fusion results are fed into a segmentation head to generate the semantic segmentation result $\hat{y}$. The segmentation head consists of a convolution layer.

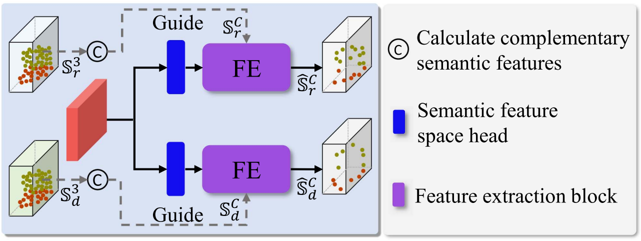

Fig. 3. Structure of the CompSemFE module. This module has two similar branches to extract complementary semantic features for each modality.

C. The CompSemFE Module

Fig. 3 shows the structure of the CompSemFE module. There are three inputs in the CompSemFE module: the fusion result of $\mathbb{R}^3$ and $\mathbb{D}^3$, the RGB semantic feature set $\mathbb{S}_r^3$, and the depth semantic feature set $\mathbb{S}_d^3$. The outputs of the CompSemFE module are RGB semantic feature complementary set $\hat{\mathbb{S}}_r^C$ and depth semantic feature complementary set $\hat{\mathbb{S}}_d^C$.

We use two semantic feature set heads to map the fusion result to the semantic feature sets. Then, we extract complementary semantic features from the semantic feature sets. Two separate feature extraction blocks are used to extract complementary semantic features for $\mathbb{S}_r^3$ and $\mathbb{S}_d^3$, respectively. The feature extraction block consists of a convolution layer, a batch normalization layer, and a ReLU layer. To ensure the extracted complementary set $\hat{\mathbb{S}}_r^C$ ($\hat{\mathbb{S}}_d^C$) is the complementary set of $\mathbb{S}_r^3$ ($\mathbb{S}_d^3$) in the $\mathbb{S}^3$, we calculate the ground-truth complementary set $\mathbb{S}_r^C$ ($\mathbb{S}_d^C$) of $\mathbb{S}_r^3$ ($\mathbb{S}_d^3$) in $\mathbb{S}^3$ and use it as guidance information to extract $\hat{\mathbb{S}}_r^C$ ($\hat{\mathbb{S}}_d^C$). The $\mathbb{S}_r^C$ is calculated as follows:

$$\mathbb{S}_r^C = \mathbb{S}^3 - \mathbb{S}_r^3 = \{x_i | x_i \neq y_i, x_i \in \mathbb{S}^3, y_i \in \mathbb{S}_r^3\}, \tag{5}$$

where i is the index of the element in $\mathbb{S}^3$, and $\neq$ means that two elements do not belong to the same class. $\mathbb{S}_d^C$ is calculated in the same way. The process of extracting $\hat{\mathbb{S}}_r^C$ is:

$$\hat{\mathbb{S}}_r^C = Q\Big(fe_r \left(conv_r \left((\mathbb{R}^3 + \mathbb{D}^3) \right) \right) | \mathbb{S}_r^C \Big), \tag{6}$$

where $+$ refers to the element-wise addition and $fe_r(\cdot)$ refers to the process of the feature extraction block for RGB modality. The $\hat{\mathbb{S}}_d^C$ is calculated in the same way. Finally, the output $\hat{\mathbb{S}}_r^C$ is fed into RGB stream and fused with $\mathbb{S}_r^3$. Similarly, the output $\hat{\mathbb{S}}_d^C$ is fed into the depth stream and fused with $\mathbb{S}_d^3$.

Fig. 4 shows a sample of some segmentation results mapped from $\mathbb{S}^3$, $\mathbb{S}_d^3$, $\mathbb{S}_d^C$, and $\hat{\mathbb{S}}_d^C$ during the semantic feature complementing process. The figure demonstrates an example of the process of semantic-feature complementary fusion for the depth modality. Firstly, we generate the depth semantic feature set $\mathbb{S}_d^3$ with the guidance of y_2. Secondly, we calculate the ground-truth complementary set for depth modality $\mathbb{S}_d^C$ between $\mathbb{S}^3$ and $\mathbb{S}_d^3$ with (5). The ground-truth complementary set $\mathbb{S}_d^C$ denotes the semantic feature information of the negative obstacles that cannot

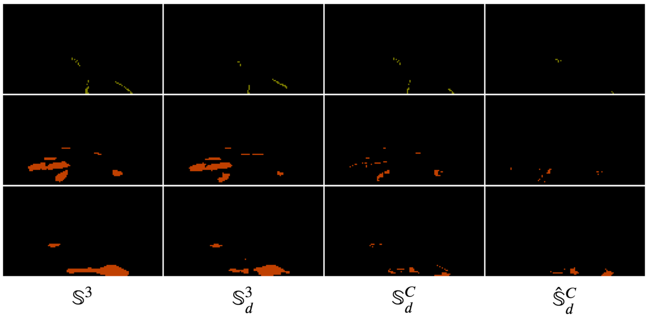

Fig. 4. Example of semantic-feature complementary fusion process for the depth modality. $\mathbb{S}^3$, $\mathbb{S}^3_d$, $\mathbb{S}^C_d$, and $\hat{\mathbb{S}}^C_d$ refer to the semantic feature set, depth semantic feature set, ground-truth complementary set, and predicted complementary set. The images are the segmentation results generated from the above sets, in which 🟥 and 🟩 represent potholes and cracks, respectively.

be extracted from the depth modality. So, in order to improve the performance of semantic segmentation, we only need to extract this semantic feature information from other modalities. Thirdly, we use the feature extraction block to extract the semantic feature information that cannot be extracted from the depth modality from the RGB modality. The $\hat{\mathbb{S}}^C_d$ is the semantic feature information extracted from the RGB modality. Finally, we fuse $\hat{\mathbb{S}}^C_d$ and $\mathbb{S}^3_d$ to increase the semantic features. From Fig. 4, we can find that $\hat{\mathbb{S}}^C_d$ has the complementary semantic features for the missing semantic feature of $\mathbb{S}^3_d$.

D. The Loss Functions

We use the cross-entropy loss $\mathcal{L}_{seg}(y, \hat{y})$ between the ground truth y and the segmentation result $\hat{y}$, which is calculated as:

$$\mathcal{L}(y, \hat{y}) = -\frac{1}{H \times W} \sum_{h=1}^{H} \sum_{w=1}^{W} \sum_{c=1}^{C} y_{(h,w,c)} \cdot \log(\hat{y}_{(h,w,c)}), \quad (7)$$

where H and W represent the height and width of the input RGB images, respectively. C represents the number of classes that need to be segmented. $\hat{y}_{h,w,c}$ refers to the probability that the network classifies the pixel at (x, y) in the image as the class c.

There are four guiding processes in the DSCF module: two processes guide the mapping process of semantic features, and two processes guide the complementary semantic feature extraction process. We also adopt cross-entropy losses to supervise the above processes. For the RGB modality, we use $\mathcal{L}_{seg}(y_2, \mathbb{S}^3_r)$ to supervise the generation of $\mathbb{S}^3_r$, which is calculated as:

$$\mathcal{L}_{seg}(y_2, \mathbb{S}^3_r) = -\frac{1}{N} \sum_{n=1}^{N} \sum_{c=1}^{C} y_{2(n,c)} \cdot \log(\mathbb{S}^3_{r(n,c)}), \quad (8)$$

where N refers to the number of elements in RGB semantic feature set $\mathbb{S}^3_r$, and $\mathbb{S}^3_{r(n,c)}$ refers to the c-th feature of n-th element in $\mathbb{S}^3_r$. The loss $\mathcal{L}_{seg}(y_2, \mathbb{S}^3_d)$ for the depth modality is calculated in the same way. We also use $\mathcal{L}_{seg}(\mathbb{S}^C_r, \hat{\mathbb{S}}^C_r)$ to supervise the extraction of complementary set $\hat{\mathbb{S}}^C_r$ for the RGB semantic feature set, which is calculated as:

$$\mathcal{L}_{seg}(\mathbb{S}^C_r, \hat{\mathbb{S}}^C_r) = -\frac{1}{N} \sum_{n=1}^{N} \sum_{m=1}^{M} \mathbb{S}^C_{r(n,m)} \cdot \log(\hat{\mathbb{S}}^C_{r(n,m)}), \quad (9)$$

where $\mathbb{S}^C_{r(n,m)}$ refers to the m-th feature of n-th element in $\mathbb{S}^C_r$. M represents the number of classes. The loss $\mathcal{L}_{seg}(\mathbb{S}^C_d, \hat{\mathbb{S}}^C_d)$ for the depth modality is calculated in the same way.

In total, we use the above 5 losses to train our PotCrackSeg. The total loss is:

$$\mathcal{L}_{total} = \mathcal{L}_{seg}(y, \hat{y}) + \sum_{i=r,d} \left(\mathcal{L}_{seg}(y_2, \mathbb{S}^3_i) + \mathcal{L}_{seg}(\mathbb{S}^C_i, \hat{\mathbb{S}}^C_i) \right). \quad (10)$$

where r and d refer to RGB and depth, respectively.

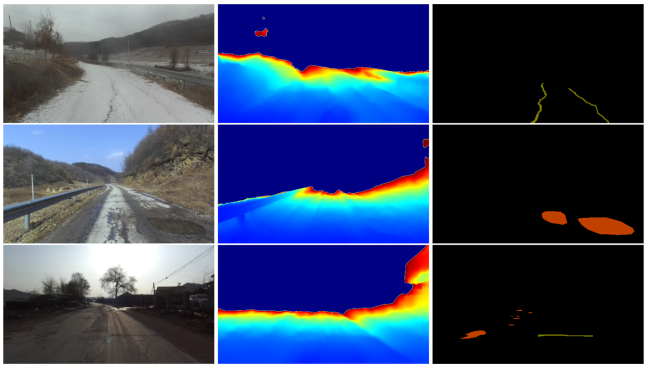

Fig. 5. Sample RGB images, depth images, and labels in our NPO++ dataset. The depth images are visualized in the *jet* color map, in which depth values increase from blue to red. 🟥 and 🟩 represent potholes and cracks.

IV. The NPO++ Dataset

We re-label the existing NPO dataset [5] to provide separate ground-truth labels for potholes and cracks. The NPO dataset is collected from urban and rural scenes with a real vehicle, and contains various weather and lighting conditions. The dataset has manually-labeled ground-truth masks for negative obstacles and positive obstacles. The negative obstacles include potholes and cracks, which are treated as one class. We separate the two classes, potholes and cracks, in this work from a total of 4,600 images. We name the upgraded dataset as NPO++. Fig. 5 shows some sample images of our dataset.

There are 4,032 images containing potholes with over 5 million pixels, and 723 images containing cracks with over 427 thousand pixels. The resolution of the images in the dataset is the same as that of the NPO dataset, that is, 288 × 512. We follow the rules of the NPO dataset to divide the images into different scenes. The dataset contains 2,661 images from urban scenes and 1,939 images from rural scenes. Moreover, there are 3,156 images from the normal-surface roads and 1,444 images from the abnormal-surface roads. To the best of our knowledge, our NPO++ dataset is the first RGB-D semantic segmentation dataset that separates road potholes and cracks.

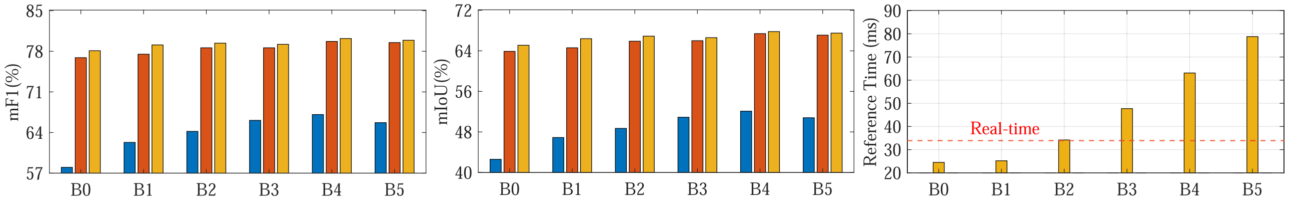

Fig. 6. The results of the ablation study on modality. 🟦, 🟧, and 🟨 represent only-depth-modality variants, only-RGB-modality variants, and multi-modal variants, respectively. We only test the inference speed of multi-modal variants. The red dashed line represents that the network is capable of real-time inference, processing at a rate of 30 frames per second.

V. Experimental Results and Discussions

A. The Dataset

We randomly split the images with crack and pothole labels from our NPO++ dataset into three sets with the ratio of 2:1:1, that is, 2,300 pairs of images for training; 1,150 pairs of images for validation; and 1,150 pairs of images for testing. We use the testing set to test the network performance. We also split the images in the testing set into different subsets according to scenes and road surfaces. In the testing set, there are 480 images from rural scenes, 670 images from urban scenes, 785 images from normal-surface roads, and 365 images from abnormal-surface roads.

B. Training Details

Our network is implemented with PyTorch and trained on a PC with two NVIDIA RTX 3060 graphics cards. The parameters of the PotCrackSeg encoder are initialized with the pre-trained weight provided from [50]. The other parameters are initialized by default in PyTorch. We use the AdamW optimizer with the initial learning rate of 6×10^{-5}, which is decayed according to the polynomial learning rate decay method with the power of 0.9. At the beginning of training, the learning rate is warmed up with the epoch of 10.

C. Ablation Study

1) Ablation on Modality: We first design a variant with only RGB modality by removing the encoder and decoder of the depth stream from PotCrackSeg. Since the input data of our proposed DSCF module are features extracted from both modalities, the DSCF module is removed along with the depth modality. Secondly, we design a variant with only depth modality by removing the encoder and decoder of the RGB stream and DSCF module from PotCrackSeg. Finally, we design several variants based on the above variants with SegFormer-B0, SegFormer-B1, SegFormer-B2, SegFormer-B3, SegFormer-B4, and SegFormer-B5 (abbreviated as B0, B1, B2, B3, B4, and B5) as encoders.

We use the metrics, mean F-score (mF1) and mean Intersection-over-Union (mIoU) [5] over both potholes and cracks, to evaluate the performance of each variant. We also measure the inference speed of the multi-modal fusion variants. The results of the variants are shown in Fig. 6. The figure shows that the results of the variants with our proposed DSCF module are better than those of the single-modal variants. This demonstrates that our proposed DSCF module can improve performance by fusing RGB and depth images containing noises. The results also demonstrate that the multi-modal variant with the B4 encoder performs the best among all the variants. Moreover, the multi-modal variant with the B2 encoder achieves real-time inference speed (30 frames per second). Based on the results, we chose two versions of PotCrackSeg: lightweight PotCrackSeg-B2 with B2 as encoder, and heavyweight PotCrackSeg-B4 with B4 as encoder.

TABLE I
The Results (%) of the Ablation Study on Fusion Strategy

Backbone	Variant	mIoU	mF1	RTX 3060	
				ms	FPS
B0	*w/o* RGB	64.5	77.6	24.2	41.3
	w/o Depth	62.6	75.9	24.3	41.2
	RGB & Depth	65.1	78.1	24.5	40.8
B1	*w/o* RGB	65.8	78.6	24.6	40.7
	w/o Depth	63.0	76.2	25.0	40.0
	RGB & Depth	66.4	79.0	25.2	39.7
B2	*w/o* RGB	66.5	79.0	34.0	29.4
	w/o Depth	64.4	77.4	34.1	29.3
	RGB & Depth	66.9	79.4	34.2	29.3
B3	*w/o* RGB	66.4	79.0	46.6	21.5
	w/o Depth	65.9	78.6	46.9	21.3
	RGB & Depth	66.6	79.2	47.7	21.0
B4	*w/o* RGB	67.5	79.8	62.7	15.9
	w/o Depth	64.4	77.5	62.6	16.0
	RGB & Depth	67.8	80.1	63.1	15.9
B5	*w/o* RGB	67.3	79.6	77.6	12.9
	w/o Depth	66.1	78.9	78.9	12.7
	RGB & Depth	67.4	79.8	78.8	12.7

w/o RGB (w/o depth) represents that no complementary set is extracted from RGB (depth) modality for depth (RGB) modality.

2) Ablation on Fusion Strategy: We conduct experiments to show the benefits brought by the dual semantic-feature complementary fusion strategy. Specifically, we design some variants with single-modal complementary fusion strategies to compare the performance with our proposed PotCrackSeg. The single-modal complementary fusion strategies are shown as:

1) *w/o* RGB: No complementary set is extracted from the RGB semantic feature set for depth modality. In contrast, the complementary set for RGB modality is extracted from the depth semantic feature set and fused into the RGB semantic feature set.
2) *w/o* Depth: No complementary set is extracted from the depth semantic feature set for RGB modality. In contrast,

TABLE II
COMPARATIVE RESULTS (%) OF DIFFERENT MODALITIES ON OUR NPO++ DATASET

Network	Years	Modality	Pothole		Crack		mIoU	mF1
			IoU	F1	IoU	F1		
RTFNet [51]	2019	RGB & Depth	66.4	79.8	10.0	18.3	38.2 (↓ 7.9)	49.0 (↓ 10.9)
		RGB	68.1	81.0	24.2	38.9	46.1	60.0
		Depth	39.6	56.8	9.1	16.7	24.4	36.7
AARTFNet [16]	2020	RGB & Depth	71.0	83.0	19.2	32.2	45.1 (↓ 4.9)	57.6 (↓ 6.7)
		RGB	69.7	82.1	30.4	46.6	50.1	64.4
		Depth	53.8	70.0	10.6	19.2	32.2	44.6
RoadSeg [21]	2020	RGB & Depth	73.9	85.0	0.0	0.0	36.9 (↓ 0.9)	42.5 (↓ 0.6)
		RGB	75.6	86.1	0.0	0.0	37.8	43.1
		Depth	62.5	76.9	0.0	0.0	31.2	38.4
TransUNet [52]	2021	RGB & Depth	71.1	83.1	40.3	57.5	55.7 (↓ 2.2)	70.3 (↓ 1.9)
		RGB	73.3	84.6	42.6	59.7	57.9	72.2
		Depth	43.4	60.5	20.0	33.4	31.7	47.0
ESANet [25]	2021	RGB & Depth	74.6	85.5	40.8	58.0	57.7 (↑ 0.7)	71.7 (↑ 0.5)
		RGB	73.6	84.8	40.4	57.6	57.0	71.2
		Depth	52.0	68.4	13.2	23.3	32.6	45.8
GMNet [26]	2021	RGB & Depth	75.2	85.8	44.4	61.5	59.8 (↓ 0.1)	73.7 (↓ 0.3)
		RGB	73.8	84.9	46.0	63.0	59.9	74.0
		Depth	56.4	72.1	24.8	39.7	40.6	55.9
MAFNet [47]	2022	RGB & Depth	69.9	82.3	26.2	41.5	48.0 (↑ 3.6)	61.9 (↑ 5.0)
		RGB	70.6	82.7	18.4	31.1	44.5	56.9
		Depth	53.4	69.6	0.1	0.1	26.7	34.9
FRNet [27]	2022	RGB & Depth	76.0	86.4	39.9	57.1	58.0 (↓ 3.0)	71.7 (↓ 3.0)
		RGB	75.6	86.1	46.3	63.3	60.9	74.7
		Depth	60.2	75.1	28.6	44.5	44.4	59.8
InconSeg [5]	2023	RGB & Depth	75.9	86.3	46.8	63.8	61.4 (↑ 3.8)	75.0 (↑ 3.1)
		RGB	72.4	84.0	42.8	60.0	57.6	72.0
		Depth	52.7	69.0	10.8	19.4	31.7	44.2
LASNet [53]	2023	RGB & Depth	57.4	72.9	42.8	59.9	50.1 (↓ 11.9)	66.4 (↓ 9.3)
		RGB	75.3	85.9	48.7	65.5	62.0	75.7
		Depth	58.8	74.1	25.1	40.2	42.0	57.1
EAEFNet [24]	2023	RGB & Depth	76.7	86.8	50.2	66.8	63.4 (↓ 0.2)	76.8 (↓ 0.1)
		RGB	77.3	87.2	50.0	66.7	63.6	76.9
		Depth	56.5	72.2	26.2	41.5	41.3	56.8
TokenFusion [28]	2022	RGB & Depth	78.1	87.7	50.6	67.2	64.4 (–)	77.5 (–)
SGFNet [30]	2023	RGB & Depth	76.6	86.7	49.1	65.8	62.8 (–)	76.3 (–)
MMSMCNet [31]	2023	RGB & Depth	66.3	79.7	29.3	45.4	47.8 (–)	62.5 (–)
PotCrackSeg-B2	Ours	RGB & Depth	*80.1*	*88.9*	*53.7*	*69.9*	**66.9** (↑ 1.0)	*79.4* (↑ 0.9)
PotCrackSeg-B4	Ours	RGB & Depth	**80.4**	**89.2**	**55.1**	**71.1**	**67.8** (↑ 0.4)	**80.1** (↑ 0.4)

The best and second-best results for each metric of each class are highlighted in bold and italic font, respectively. ↑ and ↓ respectively represent that the multi-modal results are superior and inferior to those of the single RGB modal, following the difference in parentheses.

the complementary set for depth modality is extracted from the RGB semantic feature set and fused into the depth semantic feature set.

We use the above single-modal complementary fusion strategies to design several variants with different encoders. All the variants are trained and tested with the same method and dataset. We also test the inference speed of all variants. The results are displayed in Table I. From the results, we can find that the variants with the dual semantic-feature complementary fusion strategy outperform the variants with the single-modal complementary fusion strategies. The results demonstrate that the dual semantic-feature complementary fusion strategy can better fuse the features of the two modalities and achieve better results. The inference speed for different variants shows that there is no significant increase in the inference time of the dual semantic-feature complement fusion strategy compared to the single-modal semantic feature fusion strategies.

Comparing the results of *w/o* Depth and *w/o* RGB, it can be found that the results of *w/o* RGB are better than those of *w/o* Depth. We speculate the possible reason is that the original RGB semantic feature set contains more valid information than the depth semantic feature set. It should be noted that the results of *w/o* RGB are better than those of single RGB modality, and the results of *w/o* Depth are better than those of single depth modality. This demonstrates that the semantic feature complementary fusion strategy can avoid the influence of the noises. However, the results of *w/o* Depth are inferior to those of the single RGB modality. We speculate that the network for extracting complementary features in the DSCF module is simpler, resulting in the inability to extract sufficient

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

Network	Normal						Abnormal						Urban						Rural					
	Pothole		Crack		mIoU	mF1	Pothole		Crack		mIoU	mF1	Pothole		Crack		mIoU	mF1	Pothole		Crack		mIoU	mF1
	IoU	F1	IoU	F1			IoU	F1	IoU	F1			IoU	F1	IoU	F1			IoU	F1	IoU	F1		
RTFNet [51]	62.7	77.0	7.1	13.3	34.9	45.2	74.4	85.3	12.4	22.0	43.4	53.6	62.4	76.9	6.7	12.5	34.6	44.7	72.9	84.3	11.1	19.9	42.0	52.1
AARTFNet [16]	68.3	81.2	17.0	29.1	42.7	55.1	76.6	86.8	20.9	34.6	48.8	60.7	68.7	81.4	14.4	25.2	41.5	53.3	74.8	85.6	20.7	34.3	47.7	59.9
RoadSeg [21]	71.3	83.2	0.0	0.0	35.6	41.6	79.3	88.5	0.0	0.0	39.7	44.2	71.4	83.3	0.0	0.0	35.7	41.7	77.9	87.6	0.0	0.0	39.0	43.8
TransUNet [52]	67.2	80.4	36.6	53.6	51.9	67.0	79.1	88.3	43.3	60.4	61.2	74.4	68.1	81.1	34.7	51.5	51.4	66.3	75.8	86.2	42.1	59.3	59.0	72.8
EASNet [25]	71.9	83.6	36.5	53.5	54.2	68.6	80.3	89.1	44.5	61.6	62.4	75.4	71.9	83.7	34.2	51.0	53.0	67.3	79.0	88.3	43.2	60.3	61.1	74.3
GMNet [26]	72.7	84.2	38.5	55.6	55.6	69.9	80.4	89.1	49.1	65.8	64.7	77.5	73.0	84.4	36.2	53.2	54.6	68.8	78.8	88.1	46.8	63.8	62.8	75.9
MAFNet [47]	66.0	79.5	22.1	36.2	44.0	57.8	78.0	87.6	29.6	45.7	53.8	66.7	65.9	79.5	22.0	36.0	44.0	57.8	76.2	86.5	27.6	43.3	51.9	64.9
FRNet [27]	73.6	84.8	33.3	50.0	53.5	67.4	81.0	89.5	45.0	62.0	63.0	75.8	74.0	85.1	27.1	42.6	50.5	63.8	79.2	88.4	43.6	60.7	61.4	74.6
InconSeg [5]	73.7	84.9	40.2	57.4	57.0	71.1	80.3	89.1	52.6	69.0	66.5	79.0	74.0	85.0	35.4	52.3	54.7	68.7	78.9	88.2	51.0	67.5	65.0	77.9
LASNet [53]	56.3	72.0	39.1	56.2	47.7	64.1	59.6	74.7	46.1	63.1	52.8	68.9	56.1	71.9	33.1	49.8	44.6	60.8	59.4	74.6	46.4	63.4	52.9	69.0
EAEFNet [24]	74.9	85.7	46.5	63.5	60.7	74.6	80.4	89.1	53.1	69.4	66.7	79.2	75.4	86.0	43.9	61.0	59.7	73.5	78.8	88.1	52.3	68.6	65.5	78.4
TokenFusion [28]	75.2	85.9	47.3	64.3	61.3	75.1	84.0	91.3	53.3	69.6	68.7	80.4	75.5	86.0	45.1	62.2	60.3	74.1	82.3	90.3	52.5	68.8	67.4	79.6
SGFNet [30]	74.3	85.2	45.4	62.4	59.8	73.8	81.2	89.6	52.2	68.6	66.7	79.1	74.5	85.4	42.6	59.8	58.5	72.6	79.9	88.8	51.4	67.9	65.6	78.3
MMSMCNet [31]	63.6	77.8	25.8	41.0	44.7	59.4	71.8	83.6	32.3	48.8	52.0	66.2	64.6	78.5	22.7	37.0	43.7	57.8	69.0	81.6	31.5	47.9	50.2	64.8
PotCrackSeg-B2	*77.6*	*87.4*	*50.0*	*66.6*	*63.8*	*77.0*	*85.2*	*92.0*	*56.6*	*72.3*	*70.9*	*82.1*	*78.1*	*87.7*	*50.1*	*66.8*	*64.1*	*77.2*	*83.3*	*90.9*	*54.8*	*70.8*	*69.0*	*80.8*
PotCrackSeg-B4	**78.1**	**87.7**	**52.2**	**68.6**	**65.2**	**78.2**	**85.3**	**92.0**	**57.5**	**73.0**	**71.4**	**82.5**	**78.5**	**88.0**	**51.0**	**67.6**	**64.8**	**77.8**	**83.5**	**91.0**	**56.5**	**72.2**	**70.0**	**81.6**

The best and the second best results for each metric of each class are highlighted in bold and italic font, respectively.

complementary features for the depth semantic feature set from the RGB semantic feature set.

3) Ablation on Fusion Module: We design several variants by replacing the DSCF module with the Residual-Guided Fusion (RGF) module from InconSeg [5], adaptive-mask fusion (AMF) modules from AMFNet [41], Channel-attention Fusion (CAF) module, and Dual-attention Fusion (DAF) module from MAFNet [47]. The comparative results are displayed in Table IV. We can see that the backbone with our proposed DSCF module (i.e., PotCrackSeg) achieves the best results. This demonstrates that our proposed DSCF module is more effective in tackling the issues of fusing RGB and depth images with fluctuations caused by noises.

D. Comparative Study

We compare our proposed PotCrackSeg with some well-known networks: RTFNet [51], AARTFNet [16], RoadSeg [21], TransUNet [52], EASNet [25], GMNet [26], MAFNet [47], FRNet [27], InconSeg [5], LASNet [53], EAEFNet [24], TokenFusion [28], SGFNet [30], and MMSMCNet [31]. All the networks are trained with multi-modal RGB-D images. The first 11 networks are modified to take a single modality as input to demonstrate the performance of multi-modal fusion. The networks, RTFNet, AARTFNet, RoadSeg, and TransUNet, fuse multi-modal features by simple element-wise addition or concatenation. The other networks fuse multi-modal features with fusion modules designed based on attention or designed strategies. Note that LASNet and SGFNet need binary information and boundary information of ground truth for training, so we transform ground truth labels to binary information and extract edges from ground truth labels as boundary information.

1) The Quantitative Results: The quantitative results are displayed in Table II. We can find that the multi-modal fusion results of RTFNet, AARTFNet, RoadSeg, and TransUNet are inferior to those using a single RGB modality. This illustrates that the simple element-wise addition fusion or concatenation fusion strategies could not alleviate the influence of the noises. Comparing the results of the other networks in the first ten networks, we can find that ESANet, MAFNet, and InconSeg achieve better performance when fusing RGB-D images containing the noises than only RGB images. The multi-modal fusion results of GMNet, FRNet, and EAEFNet are inferior to those using the single RGB modality. This shows that the noises in depth images influence the fusion results and even degrade the fusion results for some networks. EASNet and MAFNet adopt attention-based fusion modules to fuse features, and InconSeg uses a feature complementary strategy to fuse features. However, EAEFNet also adopts an attention-based fusion module to fuse features, but its fusion results are inferior to the results of a single RGB modality. This suggests that appropriate fusion modules and strategies can overcome the susceptibility to the noises in depth images during the fusion process.

Compared to the performance improvement of EASNet, MAFNet, and InconSeg, the improvement of our networks is not the best. Moreover, the improvement of the heavyweight PotCrackSeg-B4 is lower than that of the lightweight PotCrackSeg-B2. We conjecture that this result is mainly caused by the fact that our encoder may have extracted more valid semantic features, leaving fewer complementary semantic features to be extracted, which in turn leads to a lower performance improvement. Overall, our proposed PotCrackSeg achieves optimal results among all the compared networks, both lightweight and heavyweight.

2) The Efficiency: Fig. 8 demonstrates the efficiency of the networks. We can see that our lightweight PotCrackSeg-B2 has the fewest parameters and the third-fastest inference speed among all comparative networks. Although the speed of EASNet and FRNet is faster than PotCrackSeg-B2, PotCrackSeg-B2 outperforms them by around 10% in terms of segmentation performance. Our heavyweight PotCrackSeg-B4 is also in the middle region among all the networks in terms of parameters and inference speed. The results demonstrate that our PotCrackSeg has significant advantages in terms of speed and the amount of

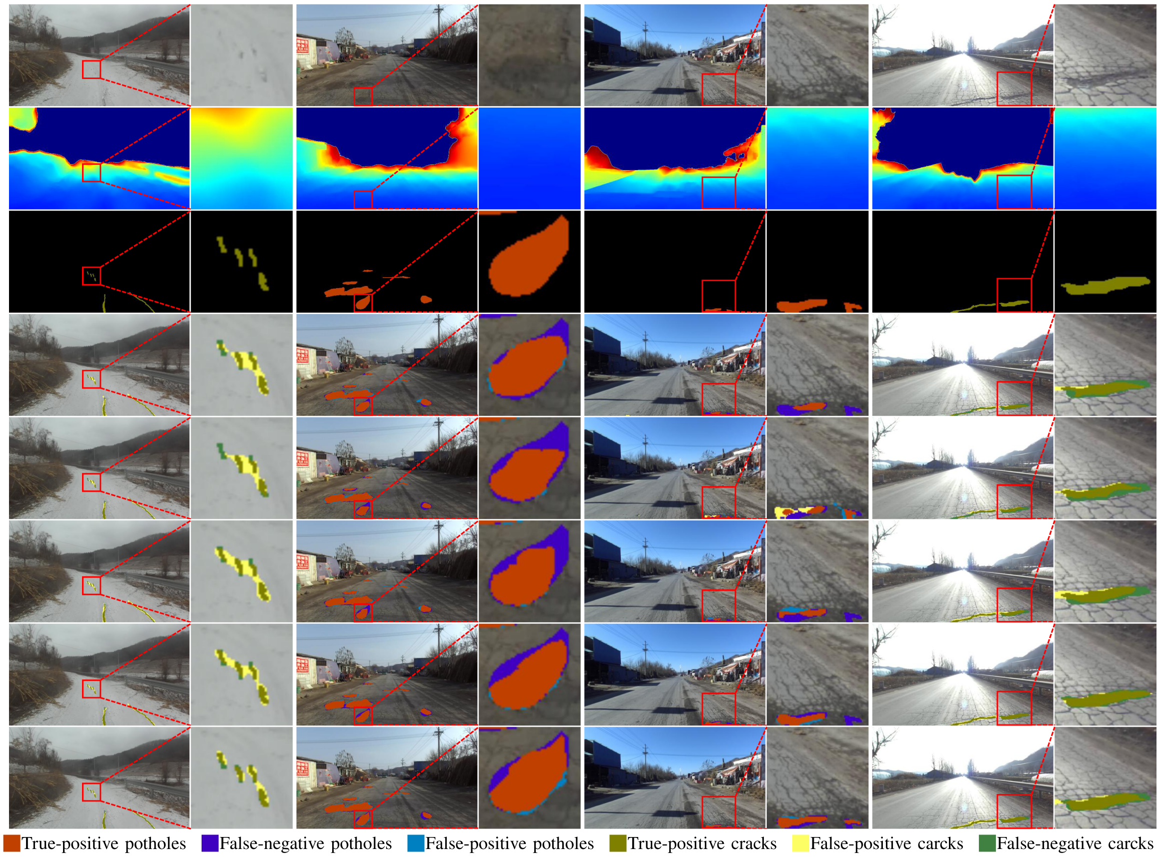

Fig. 7. Sample qualitative results for the multi-modal networks. The odd and even columns represent the complete images and enlarged views of specific regions within the complete image, respectively. The 4-th row to the last row are respectively the results of EAEFNet [24], SGFNet [30], TokenFusion [28], PotCrackSeg-B2, and PotCrackSeg-B4, respectively.

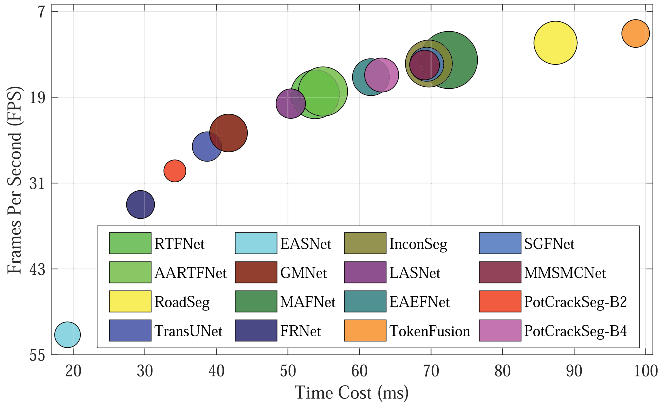

Fig. 8. Inference speed and the amount of parameters. Circles near the lower-left corner indicate faster inference speed. The radius of each circle corresponds to the amount of parameters of the network, with a larger radius indicating a larger amount of parameters. It should be noted that the amount of parameters of all networks has been normalized.

network parameters. Combined with the segmentation results, our PotCrackSeg presents a trade-off between accuracy and speed.

3) The Results on Different Scenes: We also test the performance of the multi-modal networks in different scenes. We use all images of the training set to train the networks, and use the images in different scenes of the testing set to evaluate the

TABLE IV
RESULTS (%) OF THE ABLATION STUDY ON THE FUSION MODULE

Variant	Backbone	Fusion	Network	mIoU	mF1
PotCrackSeg+SGF	B2	SGF	InconSeg	63.5	76.8
PotCrackSeg+AMF	B2	AMF	AMFNet	66.3	79.0
PotCrackSeg+CAF	B2	CAF	MAFNet	66.6	79.2
PotCrackSeg+DAF	B2	DAF	MAFNet	66.6	79.2
PotCrackSeg (Ours)	B2	DSCF	Ours	**66.9**	**79.4**
PotCrackSeg+SGF	B4	SGF	InconSeg	62.8	76.0
PotCrackSeg+AMF	B4	AMF	AMFNet	66.9	79.3
PotCrackSeg+CAF	B4	CAF	MAFNet	67.3	79.7
PotCrackSeg+DAF	B4	DAF	MAFNet	67.4	79.8
PotCrackSeg (Ours)	B4	DSCF	Ours	**67.8**	**80.1**

The best results are highlighted in bold font.

performance. Table III displays the results, which illustrates that our network achieves optimal results in all the scenes. This shows that our proposed network can be generalized to a wide range of scenes. Comparing the results between different scenes, we find that all the networks achieve better results in abnormal-surface roads and rural scenes than those in normal-surface roads and urban scenes. We conjecture the reason is that the textures of potholes and cracks in normal-surface roads and urban scenes are similar to that of roads, making it more difficult to segment potholes and cracks. One more possible reason for segmenting potholes and cracks easier in rural scenes is that the lack of effective rehabilitation of roads in rural scenes leads to large-area potholes and cracks.

4) The Qualitative Demonstrations: Some sample qualitative results of top-5 multi-modal fusion networks in Table II are shown in Fig. 7. From the first columns, we can find that segmenting the distant discontinuous cracks is a challenge. The results show that almost all the networks mis-classify the distant discontinuous cracks into one continuous crack. In contrast, our PotCrackSeg-B4 is able to segment independent cracks more accurately. This indicates that our network has a high accuracy on the target in the details. The second and the last columns also show that our PotCrackSeg presents better accuracy at the edges compared to the other networks. The third column shows that EAEFNet, SGFNet, and TokenFusion only correctly segment a small part of the pothole. However, our PotCrackSeg correctly segments most of the potholes. The results show that our PotCrackSeg achieves the optimal performance in segmenting road potholes and cracks.

VI. Conclusion and Future Work

We proposed here a novel network, PotCrackSeg, with a dual semantic-feature complementary fusion module for the segmentation of potholes and cracks in traffic scenes. We mapped both modality features into semantic feature sets and extracted the complementary semantic features for each modality to improve the fusion performance. Our proposed network alleviated the degradation caused by the noises in the depth images. In addition, we upgraded an existing large-scale RGB-D dataset by separating potholes and cracks. The experimental results demonstrate that our PotCrackSeg fusing RGB-D images containing noises outperforms using a single modality (i.e., RGB and depth). Moreover, the results also demonstrate the superiority of PotCrackSeg in terms of both accuracy and inference speed compared with well-known networks.

Although our network achieves state-of-the-art results, there are still some limitations. For example, the structure of the DSCF module to extract complementary semantic features is simple, making it less effective to extract sufficient complementary semantic features. So, in the future, we would like to design a suitable structure to extract complementary semantic features to improve the fusion performance. In addition, we would like to test the robustness of our network on the corrupted inputs with varying noise levels.

References

[1] S. Yao et al., "Radar-camera fusion for object detection and semantic segmentation in autonomous driving: A comprehensive review," *IEEE Trans. Intell. Veh.*, vol. 9, no. 1, pp. 2094–2128, Jan. 2024.

[2] S. Gao, Q. Wang, and Y. Sun, "S2G2: Semi-supervised semantic bird-eye-view grid-map generation using a monocular camera for autonomous driving," *IEEE Robot. Automat. Lett.*, vol. 7, no. 4, pp. 11974–11981, Oct. 2022.

[3] Y. Sun, W. Zuo, P. Yun, H. Wang, and M. Liu, "FuseSeg: Semantic segmentation of urban scenes based on RGB and thermal data fusion," *IEEE Trans. Automat. Sci. Eng.*, vol. 18, no. 3, pp. 1000–1011, Jul. 2021.

[4] Y. Feng, W. Hua, and Y. Sun, "NLE-DM: Natural-language explanations for decision making of autonomous driving based on semantic scene understanding," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 9, pp. 9780–9791, Sep. 2023.

[5] Z. Feng, Y. Guo, D. Navarro-Alarcon, Y. Lyu, and Y. Sun, "InconSeg: Residual-guided fusion with inconsistent multi-modal data for negative and positive road obstacles segmentation," *IEEE Robot. Automat. Lett.*, vol. 8, no. 8, pp. 4871–4878, Aug. 2023.

[6] H. Wang, Y. Sun, and M. Liu, "Self-supervised drivable area and road anomaly segmentation using RGB-D data for robotic wheelchairs," *IEEE Robot. Automat. Lett.*, vol. 4, no. 4, pp. 4386–4393, Oct. 2019.

[7] X. Ren, M. Li, Z. Li, W. Wu, L. Bai, and W. Zhang, "Curiosity-driven attention for anomaly road obstacles segmentation in autonomous driving," *IEEE Trans. Intell. Veh.*, vol. 8, no. 3, pp. 2233–2243, Mar. 2023.

[8] Y. Sun, W. Zuo, H. Huang, P. Cai, and M. Liu, "PointMoSeg: Sparse tensor-based end-to-end moving-obstacle segmentation in 3-D lidar point clouds for autonomous driving," *IEEE Robot. Automat. Lett.*, vol. 6, no. 2, pp. 510–517, Apr. 2021.

[9] S. Masihullah, R. Garg, P. Mukherjee, and A. Ray, "Attention based coupled framework for road and pothole segmentation," in *Proc. 25th Int. Conf. Pattern Recognit.*, 2021, pp. 5812–5819.

[10] X. Sun, Y. Xie, L. Jiang, Y. Cao, and B. Liu, "DMA-Net: DeepLab with multi-scale attention for pavement crack segmentation," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 10, pp. 18392–18403, Oct. 2022.

[11] L. Fan et al., "Pavement defect detection with deep learning: A comprehensive survey," *IEEE Trans. Intell. Veh.*, early access, Oct. 19, 2023, doi: 10.1109/TIV.2023.3326136.

[12] Y.-M. Kim, Y.-G. Kim, S.-Y. Son, S.-Y. Lim, B.-Y. Choi, and D.-H. Choi, "Review of recent automated pothole-detection methods," *Appl. Sci.*, vol. 12, no. 11, 2022, Art. no. 5320.

[13] Y. Bhatia, R. Rai, V. Gupta, N. Aggarwal, and A. Akula, "Convolutional neural networks based potholes detection using thermal imaging," *J. King Saud Univ.-Comput. Inf. Sci.*, vol. 34, no. 3, pp. 578–588, 2022.

[14] T. Zhao, P. Guo, J. He, and Y. Wei, "A hierarchical scheme of road unevenness perception with LiDAR for autonomous driving comfort," *IEEE Trans. Intell. Veh.*, vol. 9, no. 1, pp. 2439–2448, Jan. 2024.

[15] R. Fan, H. Wang, Y. Wang, M. Liu, and I. Pitas, "Graph attention layer evolves semantic segmentation for road pothole detection: A benchmark and algorithms," *IEEE Trans. Image Process.*, vol. 30, pp. 8144–8154, 2021.

[16] 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.

[17] Z. Feng, Y. Guo, and Y. Sun, "CPKD: Channel and position-wise knowledge distillation for segmentation of road negative obstacles," in *Proc. IEEE 26th Int. Conf. Intell. Transp. Syst.*, 2023, pp. 3110–3115.

[18] B. Kulambayev, M. Nurlybek, G. Astaubayeva, G. Tleuberdiyeva, S. Zholdasbayev, and A. Tolep, "Real-time road surface damage detection framework based on mask R-CNN model," *Int. J. Adv. Comput. Sci. Appl.*, vol. 14, no. 9, pp. 757–765, 2023.

[19] Z. S. Hernanda, H. Mahmudah, and R. W. Sudibyo, "CNN-based hyperparameter optimization approach for road pothole and crack detection systems," in *Proc. IEEE World AI IoT Congr.*, 2022, pp. 538–543.

[20] W. Xiong, J. Liu, T. Huang, Q.-L. Han, Y. Xia, and B. Zhu, "LXL: LiDAR excluded lean 3D object detection with 4D imaging radar and camera fusion," *IEEE Trans. Intell. Veh.*, vol. 9, no. 1, pp. 79–92, Jan. 2024.

[21] R. Fan, H. Wang, P. Cai, and M. Liu, "SNE-RoadSeg: Incorporating surface normal information into semantic segmentation for accurate freespace detection," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 340–356.

[22] W. Zhou, S. Dong, M. Fang, and L. Yu, "CACFNet: Cross-modal attention cascaded fusion network for RGB-T urban scene parsing," *IEEE Trans. Intell. Veh.*, vol. 9, no. 1, pp. 1919–1929, Jan. 2024.

[23] T. Zhou, S. Ruan, P. Vera, and S. Canu, "A Tri-Attention fusion guided multi-modal segmentation network," *Pattern Recognit.*, vol. 124, 2022, Art. no. 108417.

[24] M. Liang, J. Hu, C. Bao, H. Feng, F. Deng, and T. L. Lam, "Explicit attention-enhanced fusion for RGB-thermal perception tasks," *IEEE Robot. Automat. Lett.*, vol. 8, no. 7, pp. 4060–4067, Jul. 2023.

[25] 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. Automat.*, 2021, pp. 13525–13531.

[26] W. Zhou, J. Liu, J. Lei, L. Yu, and J.-N. Hwang, "GMNet: Graded-feature multilabel-learning network for RGB-thermal urban scene semantic segmentation," *IEEE Trans. Image Process.*, vol. 30, pp. 7790–7802, 2021.

[27] W. Zhou, E. Yang, J. Lei, and L. Yu, "FRNet: Feature reconstruction network for RGB-D indoor scene parsing," *IEEE J. Sel. Topics Signal Process.*, vol. 16, no. 4, pp. 677–687, Jul. 2022.

[28] 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. 12176–12185.

[29] W. Zhou, S. Dong, J. Lei, and L. Yu, "MTANet: Multitask-aware network with hierarchical multimodal fusion for RGB-T urban scene understanding," *IEEE Trans. Intell. Veh.*, vol. 8, no. 1, pp. 48–58, Jan. 2023.

[30] Y. Wang, G. Li, and Z. Liu, "SGFNet: Semantic-guided fusion network for RGB-thermal semantic segmentation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 12, pp. 7737–7748, Dec. 2023.

[31] W. Zhou, H. Zhang, W. Yan, and W. Lin, "MMSMCNet: Modal memory sharing and morphological complementary networks for RGB-T urban scene semantic segmentation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 12, pp. 7096–7108, Dec. 2023.

[32] Y. Yang, H. Yin, A.-X. Chong, J. Wan, and Q.-Y. Liu, "SACINet: Semantic-aware cross-modal interaction network for real-time 3D object detection," *IEEE Trans. Intell. Veh.*, early access, Dec. 29, 2023, doi: 10.1109/TIV.2023.3348099.

[33] Y. Liu, J. Yao, X. Lu, R. Xie, and L. Li, "DeepCrack: A deep hierarchical feature learning architecture for crack segmentation," *Neurocomputing*, vol. 338, pp. 139–153, 2019.

[34] N. H. T. Nguyen, S. Perry, D. Bone, H. T. Le, and T. T. Nguyen, "Two-stage convolutional neural network for road crack detection and segmentation," *Expert Syst. Appl.*, vol. 186, 2021, Art. no. 115718.

[35] T. Rateke and A. V. Wangenheim, "Road surface detection and differentiation considering surface damages," *Auton. Robots*, vol. 45, pp. 299–312, 2021.

[36] T. Rateke, K. A. Justen, and A. Von Wangenheim, "Road surface classification with images captured from low-cost camera-road traversing knowledge (RTK) dataset," *Revista de Informática Teórica e Aplicada*, vol. 26, no. 3, pp. 50–64, 2019.

[37] S. Guo et al., "UDTIRI: An open-source road pothole detection benchmark suite," 2023, *arXiv:2304.08842*.

[38] R. Fan and M. Liu, "Road damage detection based on unsupervised disparity map segmentation," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 11, pp. 4906–4911, Nov. 2020.

[39] 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.

[40] D. Arya et al., "Deep learning-based road damage detection and classification for multiple countries," *Automat. Construction*, vol. 132, 2021, Art. no. 103935.

[41] Z. Feng, Y. Feng, Y. Guo, and Y. Sun, "Adaptive-mask fusion network for segmentation of drivable road and negative obstacle with untrustworthy features," in *Proc. IEEE Intell. Veh. Symp.*, 2023, pp. 1–6.

[42] C. Han, T. Ma, J. Huyan, X. Huang, and Y. Zhang, "CrackW-Net: A novel pavement crack image segmentation convolutional neural network," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 11, pp. 22135–22144, Nov. 2022.

[43] J. Chen, N. Zhao, R. Zhang, L. Chen, K. Huang, and Z. Qiu, "Refined crack detection via LECSFormer for autonomous road inspection vehicles," *IEEE Trans. Intell. Veh.*, vol. 8, no. 3, pp. 2049–2061, Mar. 2023.

[44] 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.

[45] Z. Fan, H. Lin, C. Li, J. Su, S. Bruno, and G. Loprencipe, "Use of parallel resnet for high-performance pavement crack detection and measurement," *Sustainability*, vol. 14, no. 3, 2022, Art. no. 1825.

[46] Q. Zhou, Z. Qu, and F.-r. Ju, "A lightweight network for crack detection with split exchange convolution and multi-scale features fusion," *IEEE Trans. Intell. Veh.*, vol. 8, no. 3, pp. 2296–2306, Mar. 2023.

[47] 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.

[48] I. Katsamenis et al., "Deep transformer networks for precise pothole segmentation tasks," in *Proc. 16th Int. Conf. PErvasive Technol. Related Assistive Environments Conf.*, 2023, pp. 596–602.

[49] P. Liu, J. Yuan, and S. Chen, "A road damage segmentation method for complex environment based on improved UNet," in *Proc. Int. Conf. Image Graph.*, 2023, pp. 332–343.

[50] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: Simple and efficient design for semantic segmentation with transformers," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2021, pp. 12077–12090.

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

[52] J. Chen et al., "TransUNet: Transformers make strong encoders for medical image segmentation," 2021, *arXiv:2102.04306*.

[53] G. Li, Y. Wang, Z. Liu, X. Zhang, and D. Zeng, "RGB-T semantic segmentation with location, activation, and sharpening," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 3, pp. 1223–1235, Mar. 2023.

Zhen Feng (Member, IEEE) received the B.S., M.S., and Ph.D. degrees from the Harbin Institute of Technology, Harbin, China, in 2017, 2019, and 2023, respectively, and the Ph.D. degree from The Hong Kong Polytechnic University, Hung Hom, Hong Kong, in 2024. He is currently a Postdoctoral Fellow with the Department of Mechanical Engineering, The Hong Kong Polytechnic University, Hung Hom, Hong Kong. His research interests include semantic segmentation, computer vision, autonomous driving, and deep learning.

Yanning Guo received the M.S. and Ph.D. degrees in control science and engineering from the Harbin Institute of Technology, Harbin, China, in 2008 and 2012, respectively. He is currently a Professor with the Department of Control Science and Engineering, Harbin Institute of Technology. His research interests include the fields of deep space exploration, satellite attitude control, and nonlinear control.

Yuxiang Sun (Member, IEEE) received the bachelor's degree from the Hefei University of Technology, Hefei, China, in 2009, the master's degree from the University of Science and Technology of China, Hefei, China, in 2012, and the Ph.D. degree from The Chinese University of Hong Kong, Shatin, Hong Kong, in 2017. He is currently an Assistant Professor with the Department of Mechanical Engineering, City University of Hong Kong, Hong Kong. His research interests include robotics and AI, autonomous driving, autonomous systems, mobile robots, robotic perception and control, and autonomous navigation. He is also an Associate Editor for IEEE TRANSACTIONS ON INTELLIGENT VEHICLES, IEEE ROBOTICS AND AUTOMATION LETTERS, IEEE International Conference on Robotics and Automation, and IEEE/RSJ International Conference on Intelligent Robots and Systems.