Dynamic Fusion Module Evolves Drivable Area and Road Anomaly Detection: A Benchmark and Algorithms

Hengli Wang, *Graduate Student Member, IEEE*, Rui Fan, *Member, IEEE*, Yuxiang Sun, *Member, IEEE*, and Ming Liu, *Senior Member, IEEE*

***Abstract*—Joint detection of drivable areas and road anomalies is very important for mobile robots. Recently, many semantic segmentation approaches based on convolutional neural networks (CNNs) have been proposed for pixelwise drivable area and road anomaly detection. In addition, some benchmark datasets, such as KITTI and Cityscapes, have been widely used. However, the existing benchmarks are mostly designed for self-driving cars. There lacks a benchmark for ground mobile robots, such as robotic wheelchairs. Therefore, in this article, we first build a drivable area and road anomaly detection benchmark for ground mobile robots, evaluating existing state-of-the-art (SOTA) single-modal and data-fusion semantic segmentation CNNs using six modalities of visual features. Furthermore, we propose a novel module, referred to as the dynamic fusion module (DFM), which can be easily deployed in existing data-fusion networks to fuse different types of visual features effectively and efficiently. The experimental results show that the transformed disparity image is the most informative visual feature and the proposed DFM-RTFNet outperforms the SOTAs. In addition, our DFM-RTFNet achieves competitive performance on the KITTI road benchmark.**

***Index Terms*—Deep learning in robotics and automation, dynamic fusion, mobile robots, semantic scene understanding.**

Manuscript received December 9, 2020; accepted February 17, 2021. This work was supported in part by the National Natural Science Foundation of China under Grant U1713211; in part by the Hong Kong RGC Collaborative Research Fund under Grant C4063-18G; in part by the HKUST-SJTU Joint Research Collaboration Fund under Project SJTU20EG03; in part by the Young Scientists Fund of the National Natural Science Foundation of China under Grant 62003286; and in part by the HK PolyU Start-up Fund under Grant P0034801. This article was recommended by Associate Editor Y. Yuan. *(Hengli Wang and Rui Fan contributed equally to this work.) (Corresponding author: Ming Liu.)*

Hengli Wang and Ming Liu are with the Department of Electronic and Computer Engineering, The Hong Kong University of Science and Technology, Hong Kong, SAR, China (e-mail: hwangdf@connect.ust.hk; eelium@ust.hk).

Rui Fan is with the Department of Computer Science and Engineering, University of California San Diego, La Jolla, CA 92093 USA, and also with the Department of Ophthalmology, University of California San Diego, La Jolla, CA 92093 USA (e-mail: rui.fan@ieee.org).

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

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCYB.2021.3064089.

Digital Object Identifier 10.1109/TCYB.2021.3064089

I. Introduction

MOBILE robots are becoming increasingly popular in our daily life thanks to their benefits. Self-driving cars, for example, can greatly reduce traffic accidents [1]–[3]. Ground mobile robots, such as robotic wheelchairs and sweeping robots, can significantly improve people's comfort and life quality [4]–[6]. Visual environmental perception and autonomous navigation are two fundamental components in mobile robots. The former analyzes the input sensory data and outputs environmental perception results, with which the latter enables the robot to autonomously move from its current position to its destination [7]–[9]. Among all visual environmental perception tasks for mobile robots, the joint detection of drivable areas and road anomalies at the pixel level is a crucial one. Accurate and efficient drivable area and road anomaly detection can help avoid accidents for such vehicles. Note that in this article, a drivable area refers to a region in which mobile robots can navigate safely, while a road anomaly refers to a region with a height difference from the surface of the drivable area.

With the rapid development of deep-learning technologies, many semantic segmentation approaches based on convolutional neural networks (CNNs) have been proposed, and these methods can be used for drivable area and road anomaly detection. For example, Chen *et al.* [10] proposed DeepLabv3+, which combines the spatial pyramid pooling (SPP) module and the encoder–decoder architecture to generate semantic predictions. Recently, data-fusion networks have been proposed to improve the detection performance by extracting and fusing two different types of visual features. Specifically, Wang and Neumann [11] proposed a depth-aware operation to fuse RGB and depth images. In addition, Zhang *et al.* [12] fused RGB images with elevation maps. All of these fusion strategies demonstrated superior performance using multimodal data.

A. Motivation and Novel Contributions

Different types of mobile robots usually work in different environments, so they usually have different focuses on environmental perception. Self-driving cars, for example, focus mostly on cars and pedestrians because they run on roadways. Differently, ground mobile robots need to pay more attention to small obstacles, such as stones and tin cans,

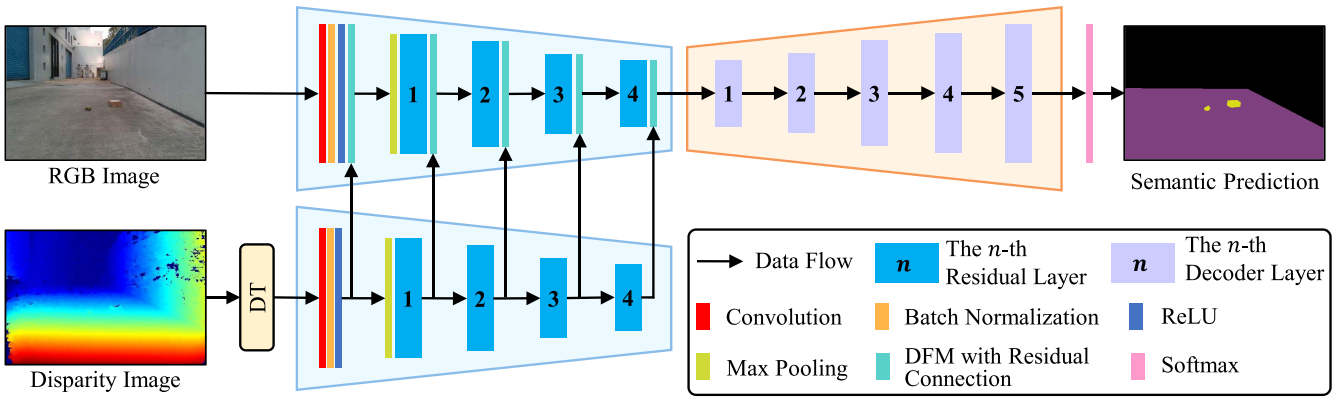


Fig. 1. Illustration of our drivable area and road anomaly detection framework. It consists of disparity transformation (DT), two encoders (the blue trapezoids), and one decoder (the orange trapezoid). DT first transforms the input disparity image. Then, our DFM dynamically fuses two different modalities of features in a multiscale fashion. Finally, the fused feature is processed by five decoder layers and a softmax layer to output the detection result.

since they usually work in indoor environments or on outdoor sidewalks. Unfortunately, current public semantic segmentation benchmarks, such as KITTI [13] and Cityscapes [14], focus mainly on driving scenarios. Other scenarios, such as robotic wheelchairs running on sidewalks, are not included. Therefore, in this article, we first build a drivable area and road anomaly detection benchmark for ground mobile robots using our previously published ground mobile robot perception (GMRP) dataset[1] [5], on which the performances of state-of-the-art (SOTA) single-modal and data-fusion networks using different types of training data are compared in detail.

Recently, incorporating different modalities of visual features into semantic segmentation has become a promising research direction that deserves more attention [15]. The visual features, such as depth and elevation, can greatly improve the detection performance. Our recent work [16] introduced a new type of visual feature, referred to as the transformed disparity image, in which the values of drivable areas become similar but the value differences between drivable areas and road anomalies/damages become significant. This can help discriminate drivable areas and road anomalies.

Moreover, the existing data-fusion networks typically fuse two different modalities of data by performing simple elementwise addition or feature concatenation. However, we can improve the detection performance by using more effective fusion operations. Inspired by the dynamic filtering network (DFN) [17], we propose a novel data-fusion module, called the dynamic fusion module (DFM), as shown in Fig. 1, which can be easily deployed in the existing data-fusion semantic segmentation networks to dynamically generate the fused feature representations using content-dependent and spatially variant kernels.

In the experiments, we first compare our proposed network, referred to as DFM-RTFNet, with ten SOTA single-modal networks and four SOTA data-fusion networks, using six different types of training data: 1) RGB images; 2) disparity images; 3) normal images; 4) HHA images; 5) elevation maps; and 6) transformed disparity images. The experimental results demonstrate that the transformed disparity image is the most informative visual feature, and it can greatly improve the drivable area and road anomaly detection performance. Moreover, our DFM-RTFNet achieves the best overall performance. We further evaluate our DFM-RTFNet on the KITTI benchmark to validate its effectiveness for self-driving cars, and the experimental results illustrate that our DFM-RTFNet achieves competitive performance on the KITTI road benchmark [18].

[1]https://github.com/hlwang1124/GMRPD

B. Article Outline

The remainder of this article is organized as follows. Section II reviews related work. Section III introduces our previously published disparity transformation algorithm used in this article. In Section IV, we introduce our DFM-RTFNet. Section V presents the experimental results and compares our framework with other SOTA approaches. Finally, we conclude this article in Section VI.

II. Related Work

In this section, we first overview some selected SOTA single-modal and data-fusion networks for semantic segmentation. We also compare these networks with our DFM-RTFNet in Sections V-C and V-E. Second, we introduce several visual features acquired from 3-D geometry information. Finally, we briefly review the dynamic filtering techniques.

A. SOTA Networks for Semantic Segmentation

FCN [19] was the first end-to-end single-modal approach for semantic segmentation. It employs an image classification network for feature extraction with the fully connected layers removed. There are three main FCN variants: 1) FCN-32s; 2) FCN-16s; and 3) FCN-8s, and we use FCN-8s in our experiments.

SegNet [20] was the first to present the encoder–decoder architecture, which is widely used in current networks. It consists of an encoder network, a corresponding decoder network, and a final pixelwise classification layer. U-Net [21] is designed based on the concept of the encoder–decoder architecture and adds skip connections between the encoder and decoder to help restore the location of small objects.

PSPNet [22] was the first to employ a pyramid pooling module to extract useful context information for better performance. DeepLabv3+ [10] follows this idea and employs depthwise separable convolution to both atrous SPP (ASPP) and the decoder module. DenseASPP [23] further connects a set of atrous convolutional layers in a dense way.

To further improve the performance, UPerNet [24] tries to identify many visual concepts, such as objects and textures in parallel; DUpsampling [25] exploits a data-dependent decoder that considers the correlation among the prediction of each pixel; and GSCNN [26] utilizes a novel architecture consisting of a shape branch and a regular branch, which can help each other focus on the relevant boundary information. Moreover, ESPNet [27] decomposes the standard convolution layer to save memory and computation cost.

As previously mentioned, data-fusion networks have been proposed to improve detection performance. Such architectures generally use two different types of visual features to learn informative representations. Specifically, FuseNet [28] adopts the encoder–decoder architecture and employs elementwise addition to fuse the feature maps of the RGB branch and depth branch. Different from FuseNet, the depth-aware CNN [11] introduces two novel operations: 1) depth-aware convolution and 2) depth-aware average pooling. These operations can leverage depth similarity between pixels to incorporate geometric information into the CNN. RTFNet [29] was developed to enhance the performance of semantic segmentation using RGB images and thermal images. It also adopts the encoder–decoder design concept and elementwise addition fusion strategy. Moreover, MFNet [30] was developed to fuse RGB images and thermal images for real-time operation.

However, these data-fusion networks often fuse two different types of information by performing simple elementwise addition or feature concatenation. We think that it is difficult to fully exploit two different modalities of data in such a simple fusion way. Different from the previous work, our proposed DFM can dynamically generate the fused feature representations using content-dependent and spatially variant kernels, which can significantly improve the detection performance.

B. Visual Features Acquired From 3-D Geometry Information

Many researchers have proposed visual features computed using 3-D geometry information to improve the drivable area and road anomaly detection performance. Specifically, Gupta *et al.* [31] combined RGB images with normal information for multiple tasks, including contour classification and semantic segmentation. Zhang *et al.* [12] fused RGB images with elevation maps to improve the semantic segmentation performance. In addition, HHA images were developed to act as complementary information for RGB images [28]. An HHA image has three channels: 1) disparity; 2) height of the pixels; and 3) the angle between the normals and the gravity vector based on the estimated ground floor. In [16] and [32], we proposed a novel visual feature, referred to as the transformed disparity image, in which the drivable areas and road anomalies are highly distinguishable. The corresponding performance comparison is presented in Sections V-C and V-D.

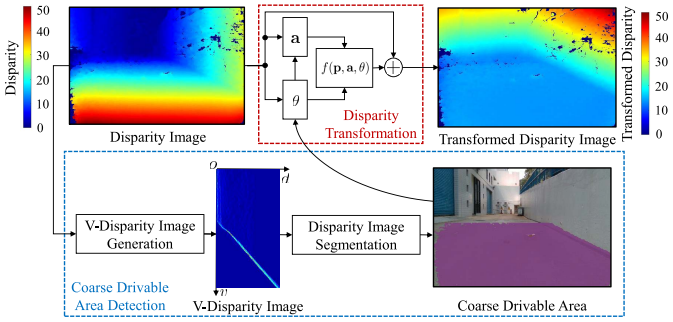


Fig. 2. Illustration of transformed disparity image generation and coarse drivable area detection.

C. Dynamic Filtering Techniques

The DFN [17] initially implemented the concept of dynamic filtering for the video prediction task, where the filters are dynamically generated based on one frame to process another frame. Recently, several extensions of the DFN have been proposed. For example, Simonovsky and Komodakis [33] extended the DFN for graph classification. Wu *et al.* [34] developed an extension of the DFN by dynamically generating weights to enlarge receptive fields for flow estimation. Our proposed DFM can also be regarded as an extension of the DFN for data-fusion semantic segmentation. We adopt the same philosophy as DFN and make specific adjustments to save GPU memory for multiscale feature fusion, which has not been studied by previous papers.

III. Disparity Transformation

DT [35] aims at transforming a disparity image into a quasi bird's eye view, so that the pixels of drivable areas possess similar values, while they differ greatly from those of the road damages/anomalies [36]. The expression of DT is as follows:

$$\mathbf{D}_{\mathrm{t}} = \mathbf{D}_{\mathrm{o}} - f(\mathbf{p}, \mathbf{a}, \theta) + \delta \tag{1}$$

where $\mathbf{D}_{\mathrm{o}}$ is the original disparity image; $\mathbf{D}_{\mathrm{t}}$ is the transformed disparity image; f is a nonlinear function representing the disparities in the drivable area; $\mathbf{p} = (u; v)$ is an image pixel; $\mathbf{a} = (a_0; a_1)$ stores the road profile model coefficients; θ is the stereo rig roll angle; and δ is a constant set to ensure that the transformed disparity values are non-negative. θ can be obtained by minimizing [37]

$$E(\theta) = \mathbf{d}^{\top}\mathbf{d} - \mathbf{d}^{\top}\mathbf{T}(\theta)\Big(\mathbf{T}(\theta)^{\top}\mathbf{T}(\theta)\Big)^{-1}\mathbf{T}(\theta)^{\top}\mathbf{d} \tag{2}$$

where $\mathbf{T} = [\mathbf{1}_k, \mathbf{v}\cos\theta - \mathbf{u}\sin\theta]$; $\mathbf{d} = (d_1; \ldots; d_k)$ is a k-entry vector of disparities; $\mathbf{u} = (u_1; \ldots; u_k)$ is a k-entry vector of horizontal coordinates; and $\mathbf{v} = (v_1; \ldots; v_k)$ denotes a k-entry vector of vertical coordinates. $\mathbf{a}$ can be estimated using [35]

$$\mathbf{a}(\theta) = \Big(\mathbf{T}(\theta)^{\top}\mathbf{T}(\theta)\Big)^{-1}\mathbf{T}(\theta)^{\top}\mathbf{d}. \tag{3}$$

In this article, we first utilize the road segmentation approach proposed in [38] to detect a coarse drivable area through v-disparity image analysis and disparity image segmentation. The disparities in the detected coarse drivable area are then used to estimate $\mathbf{a}$ and θ for disparity image transformation, as shown in Fig. 2.

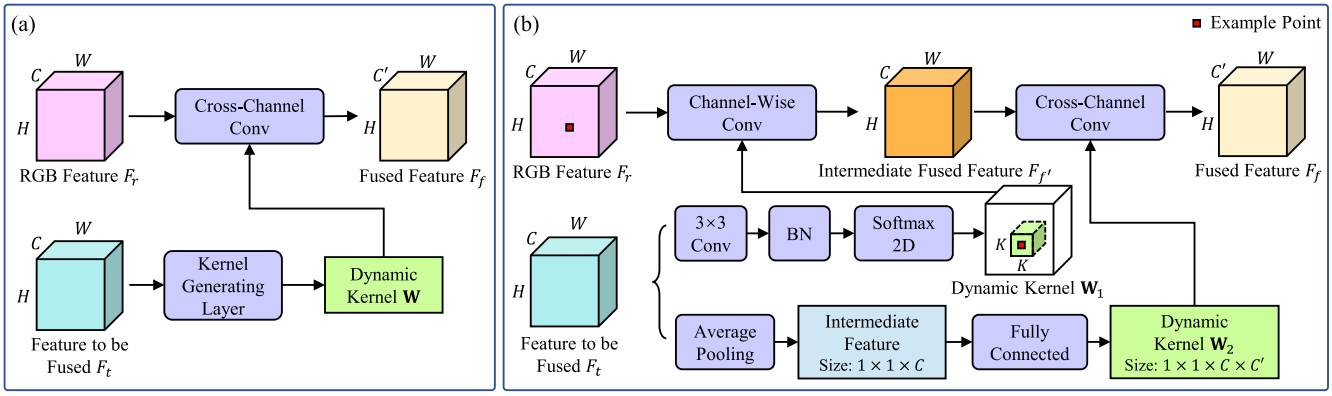


Fig. 3. Illustration of the basic concept of dynamic fusion and our proposed DFM. (a) Basic concept of dynamic fusion is to produce a dynamic kernel $\mathbf{W}$ based on the feature to be fused F_t, and then apply $\mathbf{W}$ on the RGB feature F_r to output the fused feature F_f. (b) Different from (a), our DFM adopts a two-stage convolution factorization process to save memory and computational resources.

IV. Dynamic Fusion Module

In this section, we first introduce our proposed DFM, which can generate fused features effectively and efficiently, as illustrated in Fig. 3. Then, we explain how to employ our DFM in data-fusion networks for semantic segmentation and introduce our DFM-RTFNet, as shown in Fig. 1.

Given an RGB feature F_r of size $H \times W \times C$ and a feature to be fused F_t of the same size, the basic concept of dynamic fusion is to produce a dynamic kernel $\mathbf{W}$ based on F_t and then apply $\mathbf{W}$ on F_r to output the fused feature F_f of $H \times W \times C'$, as shown in Fig. 3(a). This process can be regarded as an extension of the DFN [17] for data-fusion semantic segmentation, and has the following formulation:

$$F_f = \mathbf{W}(F_t; \Omega) \otimes F_r \tag{4}$$

where Ω denotes the parameters of the kernel generating layer shown in Fig. 3(a); and $\otimes$ denotes cross-channel convolution. The dynamic kernel $\mathbf{W}$ has two properties: 1) content dependence and 2) spatial variance. The former means that $\mathbf{W}$ is based on the feature to be fused F_t, while the latter means that different kernels are employed to different spatial positions of the RGB feature F_r. These novel properties enable the network to apply appropriate kernels to different image regions, which can generate more effective fused features F_f and thus improve the overall detection performance.

However, generating and applying the dynamic kernel in such a simple way could consume much memory and computational resources, which makes this idea hard to deploy in practice. To address this problem, we adopt the philosophy of MobileNets [39] and design a two-stage convolution factorization process for our DFM, which can save significant computational resources, as shown in Fig. 3(b).

In the first stage, we generate a dynamic kernel $\mathbf{W}_1$, which is then applied on the RGB feature F_r using the channelwise convolution operation to output an intermediate fused feature $F_{f'}$. Specifically, one channel of F_r is convolved with the corresponding channel of $\mathbf{W}_1$. The first stage can be formulated as follows:

$$F_{f'} = \mathbf{W_1}(F_t; \Omega_1) \odot F_r \tag{5}$$

where Ω_1 denotes the parameters of the kernel-generating operations shown in Fig. 3(b); and $\odot$ denotes channelwise convolution. Note that these channelwise convolutions are still spatially variant. Specifically, for an example point of F_r, we take the convolution kernel of $K \times K$ size at the corresponding position in $\mathbf{W}_1$ to conduct the channelwise convolution, as illustrated in Fig. 3(b).

In the second stage, we employ an average pooling layer and a fully connected layer to generate a dynamic kernel $\mathbf{W}_2$, which is then applied on $F_{f'}$ to generate the fused feature F_f. The second stage has the following formulation:

$$F_f = \mathbf{W_2}(F_t; \Omega_2) \otimes F_{f'} \tag{6}$$

where Ω_2 denotes the parameters of the fully connected layer. This two-stage process can significantly improve the efficiency and make this idea feasible in practice.

Then, we implement our DFM-RTFNet by integrating DFM into RTFNet50 [29], as shown in Fig. 1. Specifically, we use our DFMs with residual connections to replace the original elementwise addition layers, and our DFMs dynamically fuse two different modalities of features in a multiscale fashion. The fused feature is then processed by five decoder layers and a softmax layer sequentially to output the final detection result. For details of the network architecture, we refer readers to [29]. In addition, the number of fused feature channels is identical to the number of input feature channels, that is, $C = C'$, and we set the size of the convolution kernel $K = 3$. The determination of different fusion strategies of our DFM-RTFNet will be discussed in Section V-B.

V. Experimental Results and Discussion

A. Datasets and Experimental Setups

We utilize the following datasets in our experiments to evaluate the performance of our approach.

1) *Our GMRP Dataset [5]:* Our dataset is designed for ground mobile robots and contains 3896 pairs of images with ground truth for drivable areas and road anomalies.
2) *The KITTI Road Dataset [18]:* This dataset is designed for self-driving cars and contains 289 pairs of training data with ground truth for drivable areas and 290 pairs of testing data without ground truth.

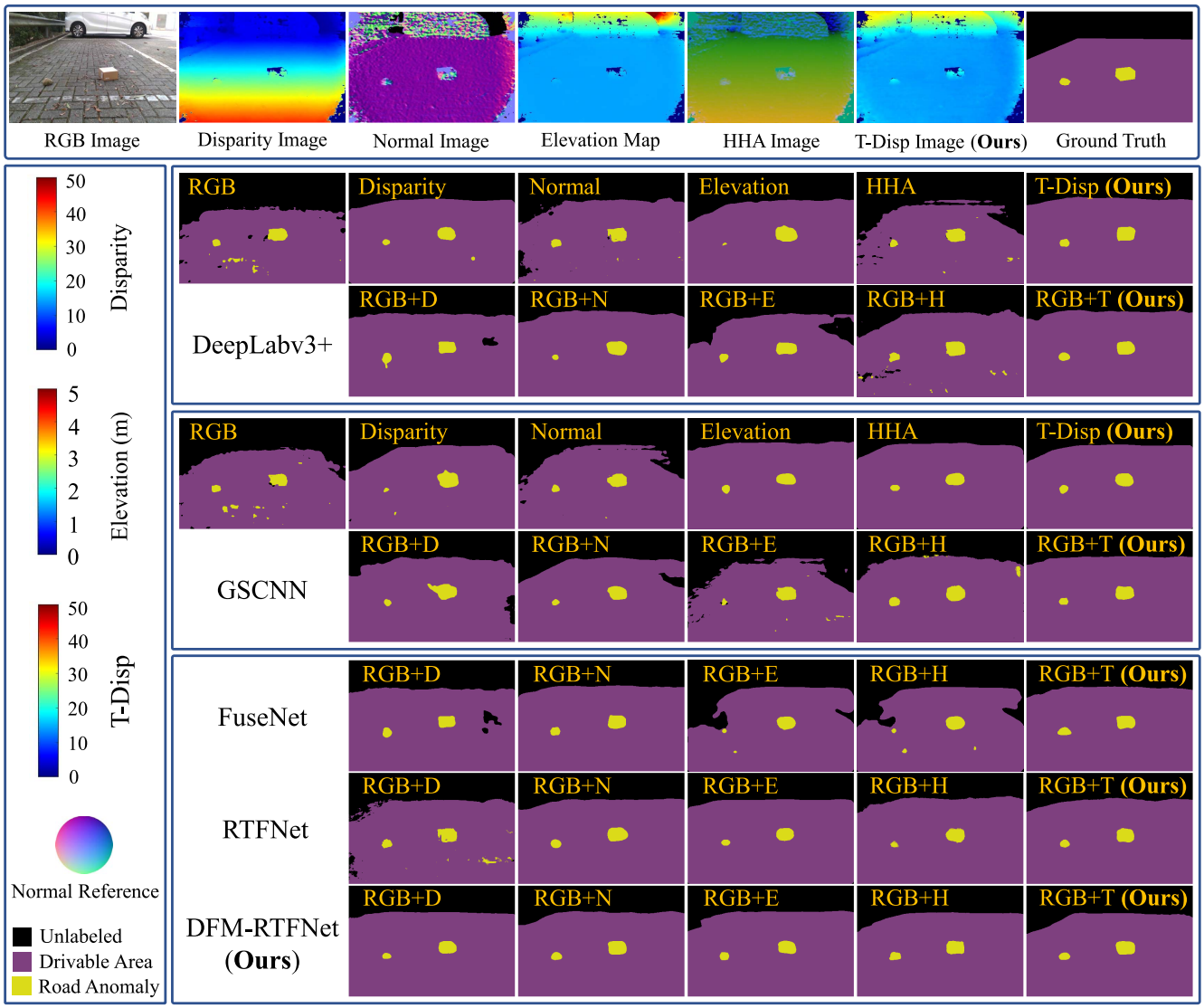


Fig. 4. Example of the experimental results on our GMRP dataset. DeepLabv3+ [10] and GSCNN [26] are single-modal networks, while FuseNet [28], RTFNet [29], and our DFM-RTFNet are data-fusion networks.

3) *The KITTI Semantic Segmentation Dataset [40]:* This dataset is also designed for self-driving cars and contains 200 pairs of training data with ground truth for scene understanding and 200 pairs of testing data without ground truth.

The total 3896 pairs of images in our GMRP dataset are split into a training, a validation, and a testing set that contain 2726, 585, and 585 pairs, respectively. The resolution of input images is downsampled to 320 × 480. We first conduct ablation studies on our GMRP dataset to: 1) select the fusion strategy of our DFM-RTFNet and 2) demonstrate the effectiveness and efficiency of our DFM, as presented in Section V-B.

Then, Section V-C presents a drivable area and road anomaly detection benchmark for ground mobile robots, which provides a detailed performance comparison of the 14 SOTA networks (ten single-modal ones and four data-fusion ones) mentioned in Section II-A and our DFM-RTFNet, with respect to six different types of training data, including our transformed disparity images. We train each single-modal network with 11 setups. Specifically, we first train each with input RGB, disparity, normal, elevation, HHA, and transformed disparity images (denoted as RGB, Disparity, Normal, Elevation, HHA, and T-Disp), respectively. Then, we train each with input concatenation of RGB and the other five types of training data separately, denoted as RGB + D, RGB + N, RGB + E, RGB + H, and RGB + T, followed by training each data-fusion network with same five setups. For the training setup, we use the stochastic gradient descent (SGD) optimizer. Moreover, we train each network until loss convergence and then select the best model according to the performance of the validation set. For the quantitative evaluations, we adopt the F-score (Fsc) and the intersection over union (IoU) for each class. We also plot the precision–recall curves and compute the average precision (AP) [41] for each class. Furthermore, we compute the mean values across all classes for the three metrics, denoted as mFsc, mIoU, and mAP.

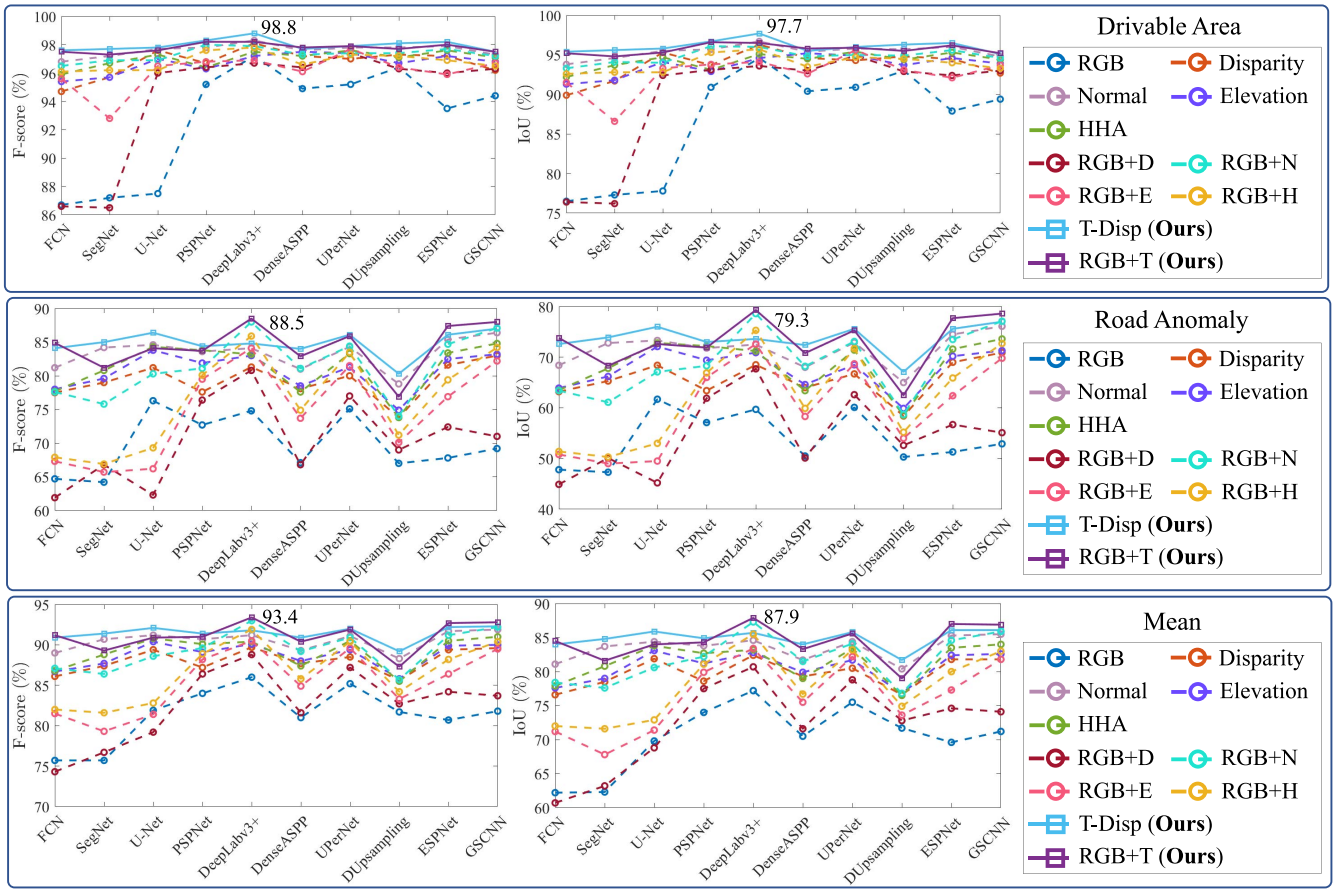


Fig. 5. Performance comparison among ten SOTA single-modal networks (FCN [19], SegNet [20], U-Net [21], PSPNet [22], DeepLabv3+ [10], DenseASPP [23], UPerNet [24], DUpsampling [25], ESPNet [27], and GSCNN [26]) with 11 training data setups on our GMRP dataset, where the best result is highlighted in each subfigure.

To better understand how our transformed disparity image improves the overall detection performance, we compare it with another two visual features that can also make the drivable areas possess similar values. In addition, we analyze the feature variation with and without our DFM to explore its internal mechanism for improving the detection performance. The experimental results are presented in Section V-D.

Finally, we conduct experiments to demonstrate the effectiveness and efficiency of our approach for self-driving cars. Since our drivable area detection task perfectly matches the KITTI road benchmark [18], we train our DFM-RTFNet using the KITTI road training data and submit it to the benchmark. However, since we focus on the detection of drivable areas and road anomalies, we do not submit the results of our approach to the KITTI semantic segmentation benchmark [40]. Instead, we merge its classes into four new classes, unlabeled, drivable area, vehicles, and pedestrians, because vehicles and pedestrians are two important anomalies for self-driving cars. We split the KITTI semantic segmentation training data into a training, a validation, and a test set that contains 100, 50, and 50 pairs of data, respectively. Then, we compare the performances between our DFM-RTFNet and four SOTA data-fusion networks with respect to six different types of training data. The experimental results are presented in Section V-E.

TABLE I
EXPERIMENTAL RESULTS OF OUR ABLATION STUDIES, WHERE (A) IS THE ORIGINAL RTFNET50 (THE BASELINE SETUP); (B) IS THE SETUP WITH FEATURE CONCATENATION; (C)–(E) ARE THREE SETUPS WITH DIFFERENT FUSION STRATEGIES OF OUR DFM-RTFNET; (F) IS THE SETUP WITH AN SOTA FUSION STRATEGY; AND (G) IS THE ORIGINAL RTFNET101. BEST RESULTS ARE BOLDED

Backbone	Fusion Strategy	mIoU (%)	Runtime (ms)	η (%/ms)
(A) RTFNet50	Addition	89.3	**24.7**	–
(B) RTFNet50	Concatenation	88.6	25.3	-1.17
(C) RTFNet50	First DFM	89.7	25.9	0.33
(D) RTFNet50	Last DFM	90.2	26.4	0.53
(E) RTFNet50	All DFMs	**92.6**	28.1	**0.97**
(F) RTFNet50	D-A Operators [11]	90.8	27.6	0.52
(G) RTFNet101	Addition	91.3	31.2	0.31

B. Ablation Study

In this section, we adopt RTFNet50 [29] with input transformed disparity images as the baseline to conduct ablation studies, and (A) of Table I shows the performance of the baseline. To compare the differences between the setups more intuitively, we introduce two new metrics: 1) the runtime of a given setup on an NVIDIA GeForce GTX 1080 Ti graphics card and 2) the ratio of the mIoU increment and runtime

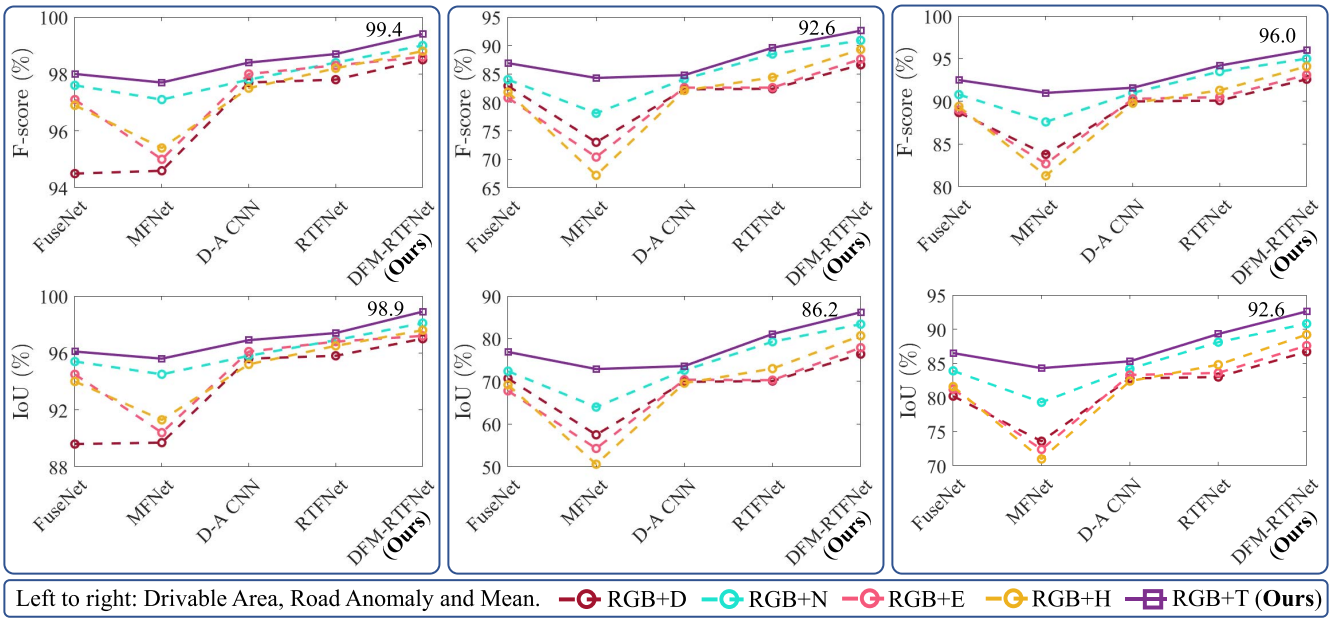


Fig. 6. Performance comparison among four data-fusion networks (FuseNet [28], MFNet [30], depth-aware CNN [11], and RTFNet [29]) and our DFM-RTFNet with five training data setups on our GMRP dataset, where D-A CNN is short for depth-aware CNN, and the best result is highlighted in each subfigure.

TABLE II
PERFORMANCE COMPARISON (%) ON OUR GMRP DATASET, WHERE AP_D AND AP_R DENOTE THE AP FOR DRIVABLE AREAS AND ROAD ANOMALIES, RESPECTIVELY. BEST RESULTS ARE BOLDED

Approach	Setup	AP_D	AP_R	mAP
DeepLabv3+ [10]	T-Disp (**Ours**)	99.71	92.45	96.08
ESPNet [27]	T-Disp (**Ours**)	99.68	91.79	95.74
GSCNN [26]	T-Disp (**Ours**)	99.36	93.61	96.49
FuseNet [28]	RGB+T (**Ours**)	99.25	93.39	96.32
RTFNet [29]	RGB+T (**Ours**)	99.70	96.27	97.99
DFM-RTFNet (**Ours**)	RGB+D	99.72	92.17	95.95
	RGB+N	99.67	97.12	98.40
	RGB+E	99.69	94.83	97.26
	RGB+H	99.61	96.13	97.87
	RGB+T (**Ours**)	**99.85**	**97.61**	**98.73**

increment between a given setup and the baseline η. Let η_i denote the η of a given setup i, and the expression of η_i is

$$\eta_i = \frac{\text{mIoU}_i - \text{mIoU}_{\text{baseline}}}{\text{Runtime}_i - \text{Runtime}_{\text{baseline}}} (\%/\text{ms}). \tag{7}$$

η is used to quantify the tradeoff between the improved performance and increased runtime of a given setup. An effective and efficient setup achieves a high η value.

We first explore different fusion strategies of our DFM-RTFNet and the corresponding performance is presented in (C)–(E) of Table I. (C) and (D) mean that we only replace the first and last elementwise addition layer, respectively, with our DFM, and (E) represents the setup shown in Fig. 1. We can observe that (C)–(E) all outperform the commonly used elementwise addition and feature concatenation strategies shown in (A) and (B) of Table I, which demonstrates that our DFM is an effective module for data fusion. Furthermore, (E) presents the best performance, and therefore, we adopt the setup illustrated in Fig. 1 in the rest of our experiments.

In addition, (F) of Table I presents the performance of the setup with an SOTA fusion strategy, depth-aware (D-A) operators [11]. We can see that our DFM outperforms it with a higher η value. Moreover, one exciting fact is that our DFM-RTFNet with the backbone of RTFNet50 [(E) of Table I] even presents a better performance than the original RTFNet101 [29] [(G) of Table I] and the runtime is much less, which demonstrates the effectiveness and efficiency of our DFM.

C. Drivable Area and Road Anomaly Detection Benchmark

In this section, we present a drivable area and road anomaly detection benchmark for ground mobile robots. Fig. 4 presents sample qualitative results, from which we can find that our transformed disparity images greatly help reduce the noises in the semantic predictions. Moreover, our DFM-RTFNet presents more accurate and robust results than all the other SOTA data-fusion networks with the same input. The corresponding quantitative performances are presented in Figs. 5 and 6, and the detailed numbers are provided in the benchmark page[2] for reference. Specifically, our DFM-RTFNet with our transformed disparity images as the input increases the mean Fsc and mean IoU by around 1.0%–21.7% and 1.8%–31.9%, respectively. We also select several networks that perform well for further performance comparison. The AP comparison is presented in Table II, and the precision–recall curves are shown in Figs. 7 and 8. We can clearly observe that our DFM-RTFNet with our transformed disparity images as input presents the best performance. This proves that our transformed disparity images and DFM can effectively improve

[2]https://sites.google.com/view/gmrb

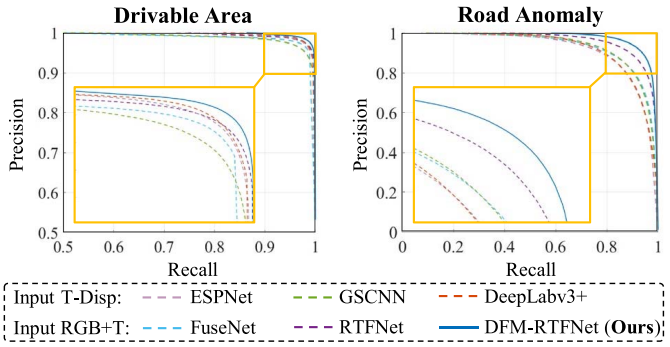


Fig. 7. Precision–recall curves of different networks on our GMRP dataset. ESPNet [27], GSCNN [26], and DeepLabv3+ [10] are single-modal networks, while FuseNet [28], RTFNet [29], and our DFM-RTFNet are data-fusion networks. The orange boxes show the enlarged area for comparison.

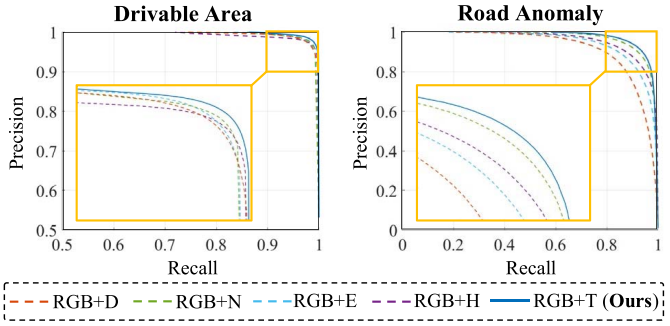


Fig. 8. Precision–recall curves of our DFM-RTFNet with five training data setups on our GMRP dataset. The orange boxes show the enlarged area for comparison.

the detection performance. In addition, we conduct more experiments to explore how our proposed approach enhances the detection performance, and the experimental results are presented in Section V-D.

D. Further Discussion

As aforementioned, our transformed disparity images can make drivable areas possess similar values, and thus improve the detection performance. To verify this, we compare our transformed disparity images with the other two visual features that have similar properties, that is: 1) normal images and 2) elevation maps. Since these three features have different scales, we introduce a dimensionless metric, the coefficient of variation c_v, defined as follows:

$$c_v = \frac{\sigma}{\mu} \tag{8}$$

where σ and μ denote the standard deviation and mean, respectively. A uniform data distribution achieves a low c_v value. Then, we compute the c_v of the three visual features on the drivable areas in our GMRP dataset. Note that for three-channel normal images, we first compute the average values across all channels and use the obtained one-channel average maps for comparison. The c_v values of normal images [42], elevation maps [43], and our transformed disparity images are 0.008, 0.009, and 0.005, respectively. We can see that our transformed disparity images achieve a much lower c_v value. The sample qualitative results are shown in Fig. 9 also verify

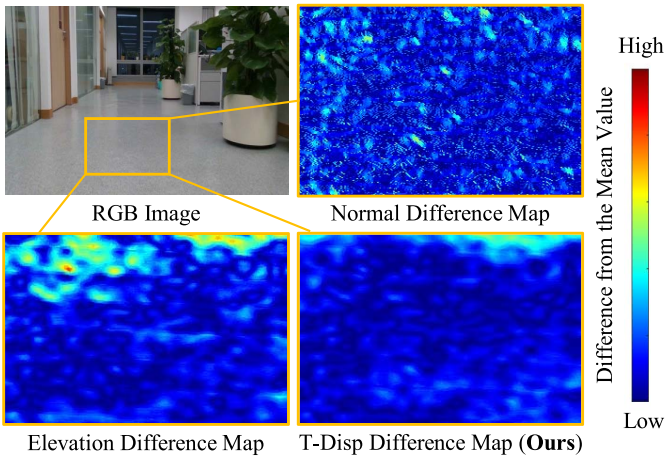


Fig. 9. Example of the drivable area consistency comparison between normal images, elevation maps, and our transformed disparity images. The orange box in the RGB image shows the enlarged area for comparison, and the other three images present the difference maps from the corresponding mean values in the enlarged area.

TABLE III
EXPERIMENTAL RESULTS (%) OF TWO SOTA DATA-FUSION NETWORKS WITH AND WITHOUT OUR DFM ON OUR GMRP DATASET. BEST RESULTS FOR EACH NETWORK ARE BOLDED

Approach	Drivable Area		Road Anomaly		Mean	
	Fsc	IoU	Fsc	IoU	Fsc	IoU
FuseNet [28]	98.0	96.1	86.9	76.9	92.5	86.5
DFM-FuseNet	**98.6**	**97.3**	**90.6**	**82.8**	**94.6**	**90.1**
RTFNet [29]	98.7	97.4	89.6	81.1	94.2	89.3
DFM-RTFNet	**99.4**	**98.9**	**92.6**	**86.2**	**96.0**	**92.6**

our conclusion that our transformed disparity images can make the drivable areas more uniform and thus benefit all networks for better detection performances.

To explore the internal mechanism of our DFM for improving the detection performance, we implement it in FuseNet [28], besides RTFNet [29], with the RGB+T setup on our GMRP dataset. The quantitative comparisons are given in Table III, where we can observe that networks with our DFM embedded generally perform better than themselves without our DFM embedded. Moreover, we visualize the mean activation maps of the features after the last layers of the encoders with and without our DFM for each network, as presented in Fig. 10. We can observe that the mean activation maps fused by our DFM can concentrate more on the drivable areas and road anomalies. We conjecture the reason is that the content-dependent and spatially variant properties make our DFM act as a weight modulation operator, which can effectively generate fused features with high activation values in the critical areas, and thus improve the detection performance.

E. Evaluations on the KITTI Benchmark

1) KITTI Road Benchmark: As previously mentioned, we submit the road estimation results of our DFM-RTFNet to the KITTI road benchmark [18]. Experimental results demonstrate that our DFM-RTFNet achieves competitive performance on

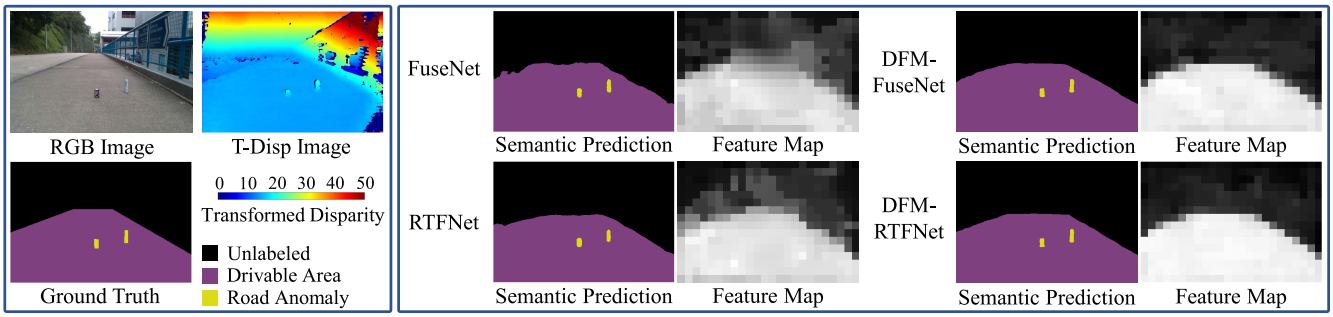


Fig. 10. Example of two SOTA data-fusion networks (FuseNet [28] and RTFNet [29]) with and without our DFM embedded, respectively. Feature maps refer to the mean activation maps of the features after the last layers of the encoders.

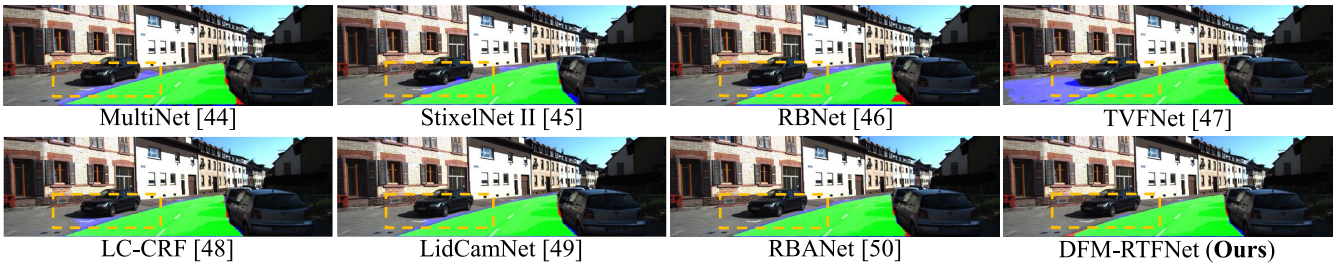


Fig. 11. Example of the testing images on the KITTI road benchmark. The true positive, false negative, and false positive pixels are shown in green, red, and blue, respectively. Significantly improved regions are marked with orange dashed boxes.

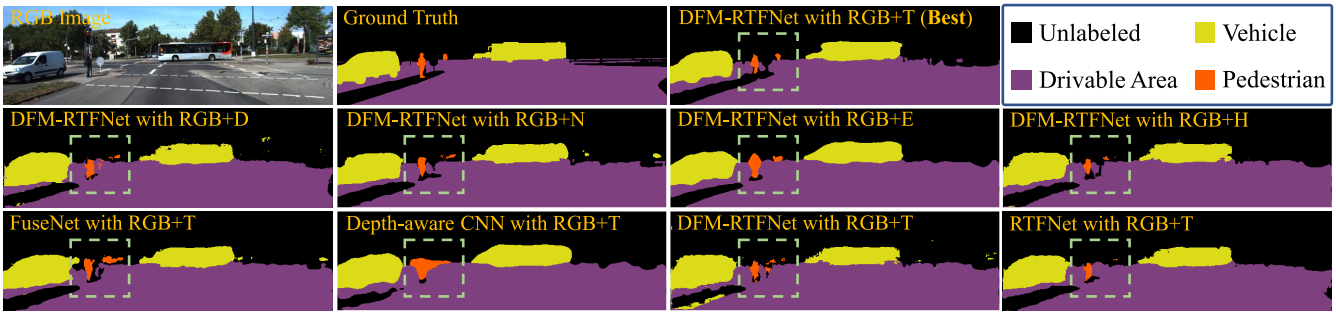


Fig. 12. Example of the experimental results on the KITTI semantic segmentation dataset. FuseNet [28], MFNet [30], Depth-aware CNN [11], RTFNet [29], and our DFM-RTFNet are all data-fusion networks. Significantly improved regions are marked with green dashed boxes.

TABLE IV
KITTI ROAD BENCHMARK[3] RESULTS, WHERE THE BEST RESULTS ARE BOLDED. NOTE THAT WE ONLY COMPARE OUR APPROACH WITH OTHER PUBLISHED APPROACHES

Approach	MaxF (%)	AP (%)	Runtime (s)
MultiNet [44]	94.88	93.71	0.17
StixelNet II [45]	94.88	87.75	1.20
RBNet [46]	94.97	91.49	0.18
TVFNet [47]	95.34	90.26	**0.04**
LC-CRF [48]	95.68	88.34	0.18
LidCamNet [49]	96.03	93.93	0.15
RBANet [50]	96.30	89.72	0.16
DFM-RTFNet (**Ours**)	**96.78**	**94.05**	0.08

the benchmark. Fig. 11 illustrates an example of the testing images, and Table IV presents the quantitative results. We can observe that our DFM-RTFNet outperforms existing approaches, which demonstrates its effectiveness and efficiency for self-driving cars.

[3]http://www.cvlibs.net/datasets/kitti/eval_road.php

TABLE V
EXPERIMENTAL RESULTS (%) OF FOUR SOTA DATA-FUSION NETWORKS AND OUR DFM-RTFNET WITH RESPECT TO DIFFERENT TRAINING DATA SETUPS ON THE KITTI SEMANTIC SEGMENTATION DATASET. BEST RESULTS ARE BOLDED

Approach	Setup	mFsc	mIoU
FuseNet [28]	RGB+T (**Ours**)	88.4	79.6
MFNet [30]	RGB+T (**Ours**)	85.5	75.2
Depth-aware CNN [11]	RGB+T (**Ours**)	87.0	77.5
RTFNet [29]	RGB+T (**Ours**)	90.3	82.7
DFM-RTFNet (**Ours**)	RGB+D	88.2	79.5
	RGB+N	90.6	83.2
	RGB+E	89.3	81.2
	RGB+H	89.8	82.1
	RGB+T (**Ours**)	**92.5**	**86.5**

2) KITTI Semantic Segmentation Dataset: Fig. 12 and Table V present the qualitative and quantitative results on the KITTI semantic segmentation dataset [40], respectively. We can see that our transformed disparity images greatly improve

the detection performance, and our DFM-RTFNet presents a better performance than all other SOTA data-fusion networks, which verifies that our DFM-RTFNet with the RGB + T setup can be deployed effectively for self-driving cars in practice.

VI. Conclusion

In this article, we conducted a comprehensive study on the drivable area and road anomaly detection for mobile robots, including building a benchmark and proposing an effective and efficient data-fusion strategy called DFM. Experimental results verified that our transformed disparity images could enable drivable areas to exhibit similar values, which could benefit networks for the drivable area and road anomaly detection. Moreover, our proposed DFM can effectively generate fused features with high activation values in critical areas using content-dependent and spatially variant kernels, and thus improve the overall detection performance. Compared with the SOTA networks, our DFM-RTFNet can produce more accurate and robust results for the drivable area and road anomaly detection. Hence, it is suitable to be implemented in mobile robots in practice. Furthermore, our DFM-RTFNet achieves competitive performance on the KITTI road benchmark. We believe that our benchmark and the data fusion idea in our proposed network will inspire future research in this area. In the future, we will develop effective and efficient data-fusion strategies to further improve the detection performance.

References

[1] R. Chai, A. Tsourdos, A. Savvaris, S. Chai, Y. Xia, and C. L. P. Chen, "Multiobjective overtaking maneuver planning for autonomous ground vehicles," *IEEE Trans. Cybern.*, early access, Mar. 5, 2020, doi: 10.1109/TCYB.2020.2973748.

[2] T. Liu *et al.*, "The role of the hercules autonomous vehicle during the covid-19 pandemic: An autonomous logistic vehicle for contactless goods transportation," *IEEE Robot. Autom. Mag.*, early access, Jan. 29, 2020, doi: 10.1109/MRA.2020.3045040.

[3] 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. Autom. Lett.*, vol. 6, no. 2, pp. 510–517, Apr. 2021.

[4] L. Li, Y.-H. Liu, T. Jiang, K. Wang, and M. Fang, "Adaptive trajectory tracking of nonholonomic mobile robots using vision-based position and velocity estimation," *IEEE Trans. Cybern.*, vol. 48, no. 2, pp. 571–582, Feb. 2018.

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

[6] H. Wang, R. Fan, Y. Sun, and M. Liu, "Applying surface normal information in drivable area and road anomaly detection for ground mobile robots," 2020. [Online]. Available: arXiv:2008.11383.

[7] Y. Sun, L. Wang, Y. Chen, and M. Liu, "Accurate lane detection with atrous convolution and spatial pyramid pooling for autonomous driving," in *Proc. IEEE Int. Conf. Robot. Biomimetics*, 2019, pp. 642–647.

[8] Y. Sun, W. Zuo, and M. Liu, "See the future: A semantic segmentation network predicting ego-vehicle trajectory with a single monocular camera," *IEEE Robot. Autom. Lett.*, vol. 5, no. 2, pp. 3066–3073, Apr. 2020.

[9] H. Wang, Y. Sun, R. Fan, and M. Liu, "S2p2: Self-supervised goal-directed path planning using rgb-d data for robotic wheelchairs," in *Proc. IEEE Int. Conf. Robot. Autom.*, to be published.

[10] 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.

[11] W. Wang and U. Neumann, "Depth-aware CNN for RGB-D segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 135–150.

[12] W. Zhang, H. Huang, M. Schmitz, X. Sun, H. Wang, and H. Mayer, "A multi-resolution fusion model incorporating color and elevation for semantic segmentation," *Int. Arch. Photogram. Remote Sens. Spat. Inf. Sci.*, vol. 42, 2017, pp. 513–517.

[13] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 3354–3361.

[14] M. Cordts *et al.*, "The cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 3213–3223.

[15] 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.

[16] R. Fan, U. Ozgunalp, B. Hosking, M. Liu, and I. Pitas, "Pothole detection based on disparity transformation and road surface modeling," *IEEE Trans. Image Process.*, vol. 29, pp. 897–908, 2020.

[17] X. Jia, B. De Brabandere, T. Tuytelaars, and L. V. Gool, "Dynamic filter networks," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2016, pp. 667–675.

[18] J. Fritsch, T. Kuehnl, and A. Geiger, "A new performance measure and evaluation benchmark for road detection algorithms," in *Proc. IEEE Conf. Intell. Transp. Syst.*, 2013, pp. 1693–1700.

[19] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3431–3440.

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

[21] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Interv.*, 2015, pp. 234–241.

[22] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2881–2890.

[23] M. Yang, K. Yu, C. Zhang, Z. Li, and K. Yang, "Denseaspp for semantic segmentation in street scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3684–3692.

[24] T. Xiao, Y. Liu, B. Zhou, Y. Jiang, and J. Sun, "Unified perceptual parsing for scene understanding," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 418–434.

[25] Z. Tian, T. He, C. Shen, and Y. Yan, "Decoders matter for semantic segmentation: Data-dependent decoding enables flexible feature aggregation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3126–3135.

[26] T. Takikawa, D. Acuna, V. Jampani, and S. Fidler, "Gated-SCNN: Gated shape cnns for semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 5228–5237.

[27] S. Mehta, M. Rastegari, A. Caspi, L. G. Shapiro, and H. Hajishirzi, "ESPNet: Efficient spatial pyramid of dilated convolutions for semantic segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 552–568.

[28] 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.

[29] 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.

[30] Q. Ha, K. Watanabe, T. Karasawa, Y. Ushiku, and T. Harada, "MFNet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2017, pp. 5108–5115.

[31] S. Gupta, P. Arbeláez, R. Girshick, and J. Malik, "Indoor scene understanding with RGB-D images: Bottom-up segmentation, object detection and semantic segmentation," *Int. J. Comput. Vis.*, vol. 112, no. 2, pp. 133–149, 2015.

[32] R. Fan, U. Ozgunalp, Y. Wang, M. Liu, and I. Pitas, "Rethinking road surface 3D reconstruction and pothole detection: From perspective transformation to disparity map segmentation," *IEEE Trans. Cybern.*, early access, Dec. 31, 2020 doi: 10.1109/TCYB.2021.3060461, 2021. [Online]. Available: 10.1109/TCYB.2021.3060461

[33] M. Simonovsky and N. Komodakis, "Dynamic edge-conditioned filters in convolutional neural networks on graphs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3693–3702.

[34] J. Wu, D. Li, Y. Yang, C. Bajaj, and X. Ji, "Dynamic filtering with large sampling field for convnets," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 185–200.

[35] 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.
[36] 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.
[37] R. Fan *et al.*, "Learning collision-free space detection from stereo images: Homography matrix brings better data augmentation," *IEEE/ASME Trans. Mechatronics*, early access, Feb. 19, 2021, doi: 10.1109/TAC.2020.2986731.
[38] U. Ozgunalp, R. Fan, X. Ai, and N. Dahnoun, "Multiple lane detection algorithm based on novel dense vanishing point estimation," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 3, pp. 621–632, Mar. 2017.
[39] A. G. Howard *et al.*, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," 2017. [Online]. Available: arXiv:1704.04861.
[40] H. A. Alhaija, S. K. Mustikovela, L. Mescheder, A. Geiger, and C. Rother, "Augmented reality meets computer vision: Efficient data generation for urban driving scenes," *Int. J. Comput. Vis.*, vol. 126, no. 9, pp. 961–972, 2018.
[41] M. Everingham, S. A. Eslami, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes challenge: A retrospective," *Int. J. Comput. Vis.*, vol. 111, no. 1, pp. 98–136, 2015.
[42] H. Badino, D. Huber, Y. Park, and T. Kanade, "Fast and accurate computation of surface normals from range images," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2011, pp. 3084–3091.
[43] S. Gupta, R. B. Girshick, P. A. Arbeláez, and J. Malik, "Learning rich features from RGB-D images for object detection and segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 345–360.
[44] M. Teichmann, M. Weber, M. Zoellner, R. Cipolla, and R. Urtasun, "Multinet: Real-time joint semantic reasoning for autonomous driving," in *Proc. IEEE Intell. Vehicles Symp.*, 2018, pp. 1013–1020.
[45] N. Garnett *et al.*, "Real-time category-based and general obstacle detection for autonomous driving," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, 2017, pp. 198–205.
[46] Z. Chen and Z. Chen, "RBNet: A deep neural network for unified road and road boundary detection," in *Proc. Int. Conf. Neural Inf. Process.*, 2017, pp. 677–687.
[47] S. Gu, Y. Zhang, J. Yang, J. M. Alvarez, and H. Kong, "Two-view fusion based convolutional neural network for urban road detection," in *IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2019, pp. 6144–6149.
[48] S. Gu, Y. Zhang, J. Tang, J. Yang, and H. Kong, "Road detection through CRF based lidar-camera fusion," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2019, pp. 3832–3838.
[49] L. Caltagirone, M. Bellone, L. Svensson, and M. Wahde, "LIDAR–Camera fusion for road detection using fully convolutional neural networks," *Robot. Auton. Syst.*, vol. 111, pp. 125–131, Jan. 2019.
[50] J.-Y. Sun, S.-W. Kim, S.-W. Lee, Y.-W. Kim, and S.-J. Ko, "Reverse and boundary attention network for road segmentation," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, 2019, pp. 876–885.

Hengli Wang (Graduate Student Member, IEEE) received the B.E. degree in mechatronics engineering from Zhejiang University, Hangzhou, China, in 2018. He is currently pursuing the Ph.D. degree with the Robotics Institute, Department of Electronic and Computer Engineering, The Hong Kong University of Science and Technology, Hong Kong, SAR, China.

His research interests include computer vision, robot navigation, and deep learning.

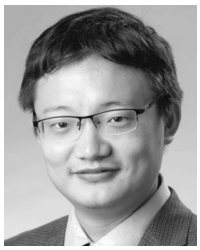
Rui Fan (Member, IEEE) received the B.Eng. degree in automation from the Harbin Institute of Technology, Harbin, China, in 2015, and the Ph.D. degree in electrical and electronic engineering from the University of Bristol, Bristol, U.K., in 2018.

From 2018 to 2020, he was a Research Associate with the Robotics Institute, The Hong Kong University of Science and Technology, Hong Kong, SAR, China. Since 2020, he has been a Postdoctoral Fellow with the Department of Computer Science and Engineering, University of California San Diego, La Jolla, CA, USA. His research interests include computer vision, image processing, high-performance computing, deep/machine learning, and autonomous driving.

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, in 2012, and the Ph.D. degree from The Chinese University of Hong Kong, Hong Kong, in 2017.

He is currently a Research Assistant Professor with the Department of Mechanical Engineering, The Hong Kong Polytechnic University, Hong Kong. He was a Research Associate with the Department of Electronic and Computer Engineering, The Hong Kong University of Science and Technology, Hong Kong, SAR, China. His current research interests include autonomous driving, deep learning, robotics and autonomous systems, and semantic scene understanding.

Dr. Sun is a recipient of the Best Paper in Robotics Award at IEEE-ROBIO 2019 and the Best Student Paper Finalist Award at IEEE-ROBIO 2015.



Ming Liu (Senior Member, IEEE) received the B.A. degree in automation from Tongji University, Shanghai, China, in 2005, and the Ph.D. degree ETH Zürich, Zürich, Switzerland, in 2013.

During his master's study with Tongji University, he stayed one year with Erlangen–Nünberg University, Erlangen, Germany, and Fraunhofer Institute IISB, Erlangen, as a Master Visiting Scholar. He is currently with the Electronic and Computer Engineering, Computer Science, and Engineering Departments, Robotics Institute, The Hong Kong University of Science and Technology, Hong Kong, SAR, China. He is also a Founding Member of Shanghai Swing Automation Company Ltd., Shanghai. He is a coordinator and is involved in NSF Projects and National 863-Hi-TechPlan Projects in China. His research interests include dynamic environment modeling, deep learning for robotics, 3-D mapping, machine learning, and visual control.

Dr. Liu was a recipient of the European Micro Aerial Vehicle Competition 2009 (second place) and two awards from the International Aerial Robot Competition 2014 as a Team Member, the Best Student Paper Award as first author for MFI 2012 (IEEE International Conference on Multisensor Fusion and Information Integration), the Best Paper Award in Information for IEEE International Conference on Information and Automation 2013 as first author, the Best Paper Award Finalists as coauthor, the Best RoboCup Paper Award for IROS 2013 (IEEE/RSJ International Conference on Intelligent Robots and Systems), the Best Conference Paper Award for IEEE-CYBER 2015, the Best Student Paper Finalist for RCAR 2015 (IEEE International Conference on Real-time Computing and Robotics), the Best Student Paper Finalist for ROBIO 2015, the Best Student Paper Award for IEEE-ICAR 2017, the Best Paper in Automation Award for IEEE-ICIA 2017, the innovation contest Chunhui Cup Winning Award in 2012 and 2013, and the Wu Weijun AI Award in 2016. He was the awardee of IEEE IROS Toshio Fukuda Young Professional Award 2018. He was the Program Chair of IEEE RCAR 2016 and the International Robotics Conference in Foshan 2017. He was the Conference Chair of ICVS 2017.