Learning Collision-Free Space Detection From Stereo Images: Homography Matrix Brings Better Data Augmentation

Rui Fan, *Member, IEEE*, Hengli Wang, *Graduate Student Member, IEEE*, Peide Cai, Jin Wu, *Member, IEEE*, Mohammud Junaid Bocus, Lei Qiao, and Ming Liu, *Senior Member, IEEE*

***Abstract*—Collision-free space detection is a critical component of autonomous vehicle perception. The state-of-the-art algorithms are typically based on supervised deep learning. Their performance is dependent on the quality and amount of labeled training data. It remains an open challenge to train deep convolutional neural networks (DCNNs) using only a small quantity of training samples. Therefore, in this article, we mainly explore an effective training data augmentation approach that can be employed to improve the overall DCNN performance, when additional images captured from different views are available. Due to the fact that the pixels in collision-free space (generally regarded as a planar surface) between two images, captured from different views, can be associated using a homography matrix, the target image can be transformed into the reference view. This provides a simple but effective way to generate training data from additional multiview images. Extensive experimental results, conducted with six state-of-the-art semantic segmentation DCNNs on three datasets, validate the effectiveness of the proposed method for enhancing collision-free space detection performance. When validated on the KITTI road benchmark, our approach provides the best results, compared with other state-of-the-art stereo vision-based collision-free space detection approaches.**

***Index Terms*—Collision-free space detection, data augmentation, homography matrix, supervised deep learning.**

Nomenclature

r, t	Pinhole cameras.
d	Disparity.
f	Camera focal length.
u	Horizontal coordinate of $\boldsymbol{p}$.
v	Vertical coordinate of $\boldsymbol{p}$.
o_u	Horizontal coordinate of $\boldsymbol{p}_o$.
o_v	Vertical coordinate of $\boldsymbol{p}_o$.
z	Depth from camera to $\boldsymbol{P}$.
$n_{x,y,z}$	x, y, and z coordinates of $\boldsymbol{n}$.
Φ	Stereo rig roll angle.
$\varkappa, \kappa$	Road disparity projection model coefficients.
$p_0 - p_5, \Delta$	Constants for Φ estimation.
c	Constant for $\varkappa$ and κ estimation.
w	Image rotation function.
m	Disparity pixel number.
E	Energy for Φ, $\varkappa$, and κ estimation.
D	Distance between r and the planar surface.
W	Image width.
T_c	Stereo rig baseline.
I	Driving scene image.
$\boldsymbol{p}$	2-D image pixel.
$\boldsymbol{p}_o$	Principal point.
$\tilde{\boldsymbol{p}}$	Homogeneous coordinates of $\boldsymbol{p}$.
$\boldsymbol{t}$	Translation vector.
$\boldsymbol{n}$	Normal vector of the planar surface.
$\boldsymbol{I}$	Identity matrix.
$\boldsymbol{P}$	3-D point in the world coordinate system.
$\boldsymbol{R}_{tr}$	Rotation matrix.
$\boldsymbol{H}_{tr}$	Homography matrix.
$\boldsymbol{K}$	Camera intrinsic matrix.

I. Introduction

THE PARADIGM in the automotive industry has shifted from high-performance cars to comfortable and safe cars in the past decade [1]. This paradigm shift has accelerated the

Manuscript received August 16, 2020; revised December 14, 2020; accepted February 6, 2021. Date of publication February 22, 2021; date of current version February 17, 2022. Recommended by Technical Editor X. Zhang and Senior Editor H. Qiao. This work was supported in part by the National Natural Science Foundation of China under Grant U1713211, in part by Collaborative Research Fund by Research Grants Council Hong Kong under Project C4063-18 G, and in part by the HKUST-SJTU Joint Research Collaboration Fund under Project SJTU20EG03, awarded to Prof. Ming Liu. *(Rui Fan and Hengli Wang are co-first authors). (Corresponding author: Ming Liu.)*

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).

Hengli Wang, Peide Cai, Jin Wu, and Ming Liu are with the Department of Electronic and Computer Engineering, Hong Kong University of Science and Technology, Hong Kong (e-mail: hwangdf@connect.ust.hk; peide.cai@connect.ust.hk; jwucp@connect.ust.hk; eelium@ust.hk).

Mohammud Junaid Bocus is with the Department of Electrical and Electronic Engineering, University of Bristol, BS8 1UB Bristol, U.K. (e-mail: junaid.bocus@bristol.ac.uk).

Lei Qiao is with the State Key Laboratory of Ocean Engineering and the School of Naval Architecture, Ocean and Civil Engineering, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: qiaolei@sjtu.edu.cn).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TMECH.2021.3061077.

Digital Object Identifier 10.1109/TMECH.2021.3061077

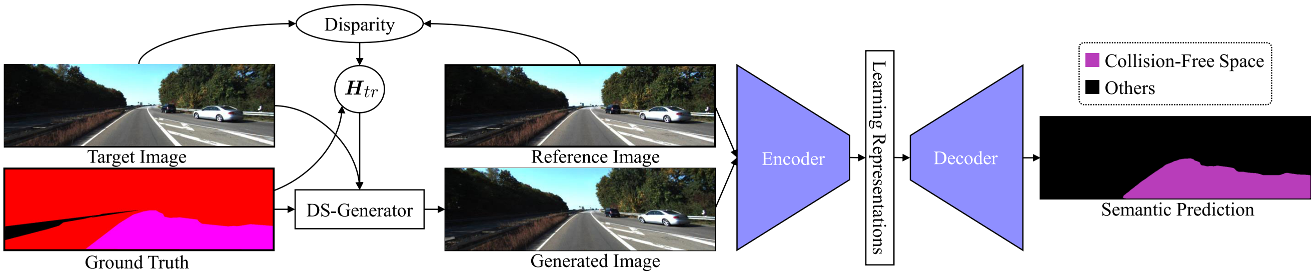

Fig. 1. Block diagram of our proposed collision-free space detection approach.

development of autonomous driving technologies, such as the Internet of vehicles [2] and advanced driver assistance systems (ADAS). In recent years, industry titans, such as Waymo, BMW, Tesla, and Volvo, have been competing with each other to commercialize autonomous vehicles [3]. However, a number of accidents occurred during experiments recently, and this has cast doubt on whether the autonomous driving technology is safe enough for deployment [4]. In this regard, the self-driving industry is now becoming more realistic. Many of them believe that the current research and development of autonomous driving technologies should still focus on the ADAS [5], [6].

Visual environment perception is a key component of the ADAS [3]. Its tasks include [7]: 1) 3-D information acquisition; 2) object detection/recognition; 3) semantic segmentation. Collision-free space detection, also referred to as occupancy grid mapping or drivable area detection, is an important task in visual environment perception [8]. Collision-free space detection approaches generally classify each pixel in the image as positive (drivable) or negative (undrivable) [9]. Such classification results are then used by other autonomous car modules, *e.g.,* trajectory prediction [10], lane departure warning [11], and obstacle avoidance [12], to ensure that the autonomous car can safely navigate in complex environments.

Recent deep convolutional neural network (DCNN)-based collision-free space detection approaches perform incredibly well [13], [14]. However, the quality and amount of training samples can greatly affect the performance of these DCNNs. In this regard, training data augmentation is generally performed to increase the diversity of the available data, without actually collecting new data. The most common way of training data augmentation is to apply different types of image transformation operations, such as reflections, rotations, and translations, to the existing data. Fortunately, for a multicamera system, such as a stereo rig, multiview images are available. However, the aforementioned image transformation operations do not consider the relationship among images captured at different view points. Therefore, jointly exploring effective training data augmentation approaches and leveraging the relationship among multiview images, especially for stereo images, has become a popular area of research that requires more attention.

The collision-free space can be considered as a planar surface. Since the 3-D points on the same planar surface between two images captured from different views can be linked by a homography matrix [15], the target image can be transformed into its reference view [16]. Hence in this article, we propose an effective driving scene generator (DS-Generator), which can produce additional RGB images for training data augmentation. The block diagram of our proposed collision-free space detection approach is shown in Fig. 1. The 3-D points on the collision-free space between the reference and target images are first used to estimate their corresponding homography matrix. The target image and the estimated homography matrix then serve as the input to our DS-Generator, and a driving scene image can be generated. Since the generated image is in the same view of the reference image, they can use the same ground truth label. To validate the effectiveness of our DS-Generator, we train six state-of-the-art semantic segmentation DCNNs on three road segmentation datasets for collision-free space detection. Extensive experiments illustrate that our DS-Generator can effectively augment training sets and all the evaluated DCNNs achieve better results for collision-free space detection. When validated on the KITTI road benchmark[1] [17], our approach provides the best results, compared with other state-of-the-art stereo vision-based collision-free space detection approaches.

The rest of this article is organized as follows. Section II provides an overview of the state-of-the-art collision-free space detection approaches. Section III introduces our DS-Generator for training data augmentation. Section IV shows the experimental results of the six state-of-the-art DCNNs and demonstrates the effectiveness of our DS-Generator for enhancing collision-free space detection. Finally, Section V concludes this article.

II. Related Work

The state-of-the-art collision-free space detection algorithms are generally grouped into two classes: 1) geometry-based; and 2) deep learning-based. The geometry-based algorithms typically formulate collision-free space with an explicit geometry model, *e.g.,* a straight line [18] or a quadratic surface [19], and find its best coefficients using optimization approaches, such as gradient descent [18] or singular value decomposition [19]. The collision-free space can then be detected by comparing the difference between the actual and modeled road surfaces [19]. Reference [20] is a typical geometry-based collision-free space detection algorithm, where the road segmentation was performed by fitting a B-spline model [21] to the road disparity projections on a 2-D disparity histogram (referred to as *v-disparity image* [22]). Similarly, [23] considered road surface

[1] [Online]. Available: www.cvlibs.net/datasets/kitti/eval_road.php

modeling as a shortest path problem and extracted the road disparity projections from the v-disparity image using Dijkstra algorithm [24]. Moreover, [19] and [25] formulated the road disparity projection modeling into a more general way by incorporating the stereo rig roll angle into the least squares fitting process, which can produce more robust results when the stereo rig baseline is not perfectly parallel to the collision-free space [25].

With recent advances in machine learning, collision-free space detection is regarded as a part of semantic driving scene segmentation, where DCNNs are proven to be the best solution. Since [26] introduced fully convolutional network (FCN), research on semantic driving scene segmentation has experienced a major boost. SegNet [27] presented the encoder–decoder architecture, which is widely utilized in current networks. The encoder network performs convolutions and max-poolings, while the decoder network uses the transferred pooling indices from the encoder to produce a sparse feature map, which is then fed to a trainable filter bank to produce a dense feature map [27]. Finally, a softmax classifier is used for the classification of each image pixel. U-Net [28] was designed based on FCN [26]. It consists of a contracting path and an expansive path [28]. The former includes convolutions, rectified linear units, and max pooling layers, while the latter combines the feature and spatial information through a sequence of upconvolutions and concatenations with the corresponding feature map from the contracting path [28].

DeepLabv3+ [29] was improved from DeepLabv1 [30], DeepLabv2 [31], and DeepLabv3 [32]. It was designed to combine the advantages of both the spatial pyramid pooling (SPP) module and the encoder–decoder architecture. It applies the depthwise separable convolution to both atrous SPP (ASPP) and the decoder module, which makes its encoder–decoder module much faster and more robust [29]. In [31], ASPP was proposed to concatenate multiple atrous-convolved features into a final feature map. However, the feature resolution is not dense enough for semantic driving scene segmentation. DenseASPP [33] was proposed to solve this problem, by connecting a set of atrous convolutional layers (ACLs) in a dense way. The ACLs in DenseASPP are organized in a cascade fashion, where the dilation rate increases layer by layer [33]. Then, DenseASPP concatenates the output of each atrous layer with the input feature map and all the outputs from lower layers. The final output of DenseASPP is a feature map generated by multiscale atrous convolutions [33]. For recent approaches with encoder–decoder architectures, the last layer of the decoder is typically a bilinear upsampling procedure for final pixelwise prediction recovery.

However, the simple bilinear upsampling has limited ability to accurately recover the pixelwise prediction, because it does not take the correlation among the prediction of each pixel into account [34]. Data-dependent upsampling (DUpsampling) [34] was designed to solve this problem, by exploiting the redundancy in the label space of semantic image segmentation and recovering the pixelwise prediction from low-resolution outputs of DCNNs. Due to the effectiveness of DUpsampling, the encoder can avoid the excessive reduction of its overall strides and this can in turn reduce the consumption of computation and memory resources dramatically [34].

Different from the aforementioned DCNNs, Gated-SCNN (GSCNN) [35] utilizes a novel two-branch architecture, which consists of a shape branch and a regular branch. Specifically, the regular branch can be any backbone architecture, and the shape branch processes the shape information in parallel to the regular branch through a set of residual blocks and gated convolutional layers. Then, GSCNN uses the higher-level activations in the regular branch to effectively help the shape branch only focus on the relevant boundary information [35]. Finally, GSCNN employs an ASPP to combine the information from the two streams in a multiscale fashion.

III. Methodology

We have two pinhole cameras r and t,[2] looking at a 3-D point $\boldsymbol{P}_i$ on a planar surface in the world coordinate system. The image pixel ${}^r\boldsymbol{p}_i = ({}^r u_i; {}^r v_i)$ of $\boldsymbol{P}_i$ captured by r and the image pixel ${}^t\boldsymbol{p}_i = ({}^t u_i; {}^t v_i)$ of $\boldsymbol{P}_i$ captured by t can be linked using [16]

$$ {}^t\tilde{\boldsymbol{p}}_i = \boldsymbol{H}_{tr}\, {}^r\tilde{\boldsymbol{p}}_i \tag{1} $$

where ${}^{r,t}\tilde{\boldsymbol{p}}$ is the homogeneous coordinates of ${}^{r,t}\boldsymbol{p}$, and the expression of the homograph matrix $\boldsymbol{H}_{tr}$ is [15]

$$ \boldsymbol{H}_{tr} = \frac{{}^r z_i}{{}^t z_i} \boldsymbol{K}_t \cdot \left(\boldsymbol{R}_{tr} - \frac{\boldsymbol{t}_{tr} \boldsymbol{n}^\top}{D} \right) \cdot \boldsymbol{K}_r^{-1} \tag{2} $$

where ${}^r z_i$ and ${}^t z_i$ are the z coordinates of $\boldsymbol{P}_i$ in the r and t camera coordinates systems, respectively; $\boldsymbol{R}_{tr}$ is the rotation matrix by which r is rotated with respect to t; $\boldsymbol{t}_{tr}$ is the translation vector from r to t; $\boldsymbol{K}_r$ and $\boldsymbol{K}_t$ are the intrinsic matrices of r and t, respectively; $\boldsymbol{n} = (n_x; n_y; n_z)$ is the normal vector of the collision-free space; and D is the distance between r and the collision-free space. For a stereo rig, ${}^r z_i = {}^t z_i$, $\boldsymbol{R}_{tr}$, $\boldsymbol{t}_{tr}$, $\boldsymbol{K}_r$, and $\boldsymbol{K}_t$ can be obtained from stereo rig calibration, $\boldsymbol{R}_{tr} = \boldsymbol{I}$, and $\boldsymbol{t}_{tr} = (Tc; 0; 0)$, where Tc is the stereo rig baseline

$$ \boldsymbol{K}_r = \boldsymbol{K}_t = \begin{bmatrix} f & 0 & o_u \\ 0 & f & o_v \\ 0 & 0 & 1 \end{bmatrix} \tag{3} $$

f is the camera focal length, and $\boldsymbol{p}_o = (o_u, o_v)$ is the principal point. (2) can, therefore, be rewritten as

$$ \boldsymbol{H}_{tr} = \begin{bmatrix} 1 - \frac{Tcn_x}{D} & -\frac{Tcn_y}{D} & \frac{o_u Tcn_x}{D} + \frac{o_v Tcn_y}{D} - \frac{fTcn_z}{D} \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}. \tag{4} $$

Equation (4) can be further written in a simplified form as follows [36]:

$$ \boldsymbol{H}_{tr} = \varkappa \begin{bmatrix} \frac{1}{\varkappa} + \sin\Phi & -\cos\Phi & -\kappa \\ 0 & 1/\varkappa & 0 \\ 0 & 0 & 1/\varkappa \end{bmatrix} \tag{5} $$

where Φ is the stereo rig roll angle, $\varkappa$ and κ are two road disparity projection model coefficients [37]. They can be estimated by

[2] r and t refer to “reference” and “target,” respectively.

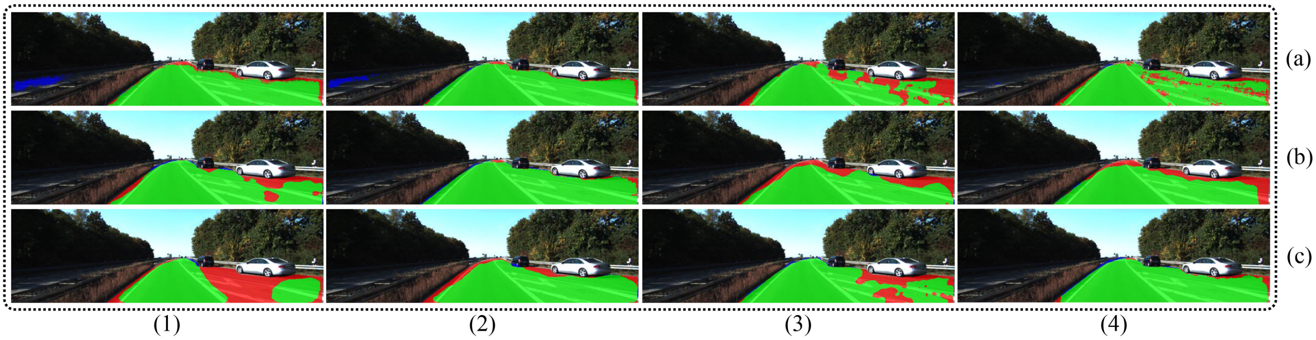

Fig. 2. Examples of the experimental results on the KITTI road dataset [17]: columns (1)-(2) on rows (a)-(c) show the experimental results of (a) SegNet [27], (b) DeepLabv3+ [29] and (c) DUpsampling [34], trained on the original and augmented training sets, respectively; columns (3) and (4) on rows (a)–(c) show the experimental results of (a) U-Net [28], (b) DenseASPP [33], and (c) GSCNN [35], trained on the original and augmented training sets, respectively. The true positive, false negative, and false positive pixels are shown in green, red, and blue, respectively.

minimizing [38]

$$E(\Phi, \varkappa, \kappa) = \sum_{i=1}^{m} \left(d_i - \varkappa \Big(w({}^{r}\boldsymbol{p}_i, \Phi) + \kappa \Big) \right)^2 \tag{6}$$

where

$$w({}^{r}\boldsymbol{p}_i, \Phi) = {}^{r}v_i \cos \Phi - {}^{r}u_i \sin \Phi. \tag{7}$$

$\min E(\Phi, \varkappa, \kappa)$ has a closed-form solution [36]

$$\Phi = \arctan \left(\frac{p_4 p_1 - p_3 p_2 + q\sqrt{\Delta}}{p_3 p_0 + p_5 p_2 - p_5 p_1 - p_4 p_0} \right) \quad s.t.\ q \in \{-1, 1\} \tag{8}$$

$$\varkappa = \frac{1}{c} \left(m \sum_{i=1}^{m} d_i w({}^{r}\boldsymbol{p}_i, \Phi) - \sum_{i=1}^{m} d_i \sum_{i=1}^{m} w({}^{r}\boldsymbol{p}_i, \Phi) \right) \tag{9}$$

$$\kappa = \frac{1}{\varkappa c} \left(\sum_{i=1}^{m} d_i \sum_{i=1}^{m} w({}^{r}\boldsymbol{p}_i, \Phi)^2 - \sum_{i=1}^{m} w({}^{r}\boldsymbol{p}_i, \Phi) \sum_{i=1}^{m} d_i w({}^{r}\boldsymbol{p}_i, \Phi) \right) \tag{10}$$

where

$$c = m \sum_{i=1}^{m} w({}^{r}\boldsymbol{p}_i, \Phi)^2 - \left(\sum_{i=1}^{m} w({}^{r}\boldsymbol{p}_i, \Phi) \right)^2. \tag{11}$$

The expressions of p_0–p_5 and Δ are given in [25]. Φ can be determined by separately replacing q in (8) with -1 and 1 and finding the minimum $\min E$ [25]. With the estimated Φ, $\varkappa$, and κ, the target image ${}^{t}I$ can be used to generate an image ${}^{g}I$ in the reference view using

$${}^{g}I(\boldsymbol{p}_i) = \begin{cases} {}^{r}I(\boldsymbol{p}_i) & \text{if } u_i - (\varkappa(w(\boldsymbol{p}_i, \Phi) + \kappa) \leq 0 \\ & \text{or } u_i - (\varkappa(w(\boldsymbol{p}_i, \Phi) + \kappa) > W \\ {}^{t}I(\boldsymbol{p}_i - (\varkappa(w({}^{r}\boldsymbol{p}_i, \Phi) + \kappa); 0)) & \text{otherwise} \end{cases} \tag{12}$$

where $\boldsymbol{p}_i$ is a 2-D pixel in the generated image ${}^{g}I$ and W is the image width. ${}^{r}I$ and ${}^{g}I$ then use the ground truth label of ${}^{r}I$ to train the DCNN.

IV. Experimental Results

A. Datasets

We conduct the experiments on the following three datasets.

1) The KITTI road dataset [17]: This dataset provides stereo image pairs, collected in real-world environments. We split it into three sets: 1) training (173 pairs of stereo images); 2) validation (58 pairs of stereo images); 3) testing (58 pairs of stereo images). The disparity information is acquired by PSMNet [40].
2) The SYNTHIA road dataset [39]: This dataset provides stereo image pairs acquired in simulation environments. We select 300 images from it and split them into three sets: training (180 pairs of stereo images), validation (60 pairs of stereo images), and testing (60 pairs of stereo images). This dataset provides the disparity ground truth.
3) Our SYN-Stereo road dataset: We publish a multiview synthetic dataset, named SYN-Stereo road dataset. This dataset is created using CARLA[3] simulator [41]. We first mount a simulated stereo rig (baseline: 1.5 m) on the top of a vehicle to capture synchronized stereo images (resolution: 640×480 pixels). The vehicle then navigates in different maps under different illumination and weather conditions, *e.g.,* clear, rainy, daytime, and sunset, for driving scene collection. We set random pedestrians including adults and children walking along the sidewalks. We also randomly set different types of vehicles, such as cars and motorcyclists, navigating in the scenarios at different speeds. The pedestrians and vehicles are all controlled by the CARLA simulator. We select 300 pairs of stereo images with corresponding disparity and semantic segmentation ground truth for collision-free space detection. We split them into three sets: 1) training (180 pairs of stereo images); 2) validation (60 pairs of stereo images); and 3) testing (60 pairs of stereo images). Our dataset is publicly available at sites.google.com/view/syn-stereo for research purposes.

[3] [Online]. Available: carla.org

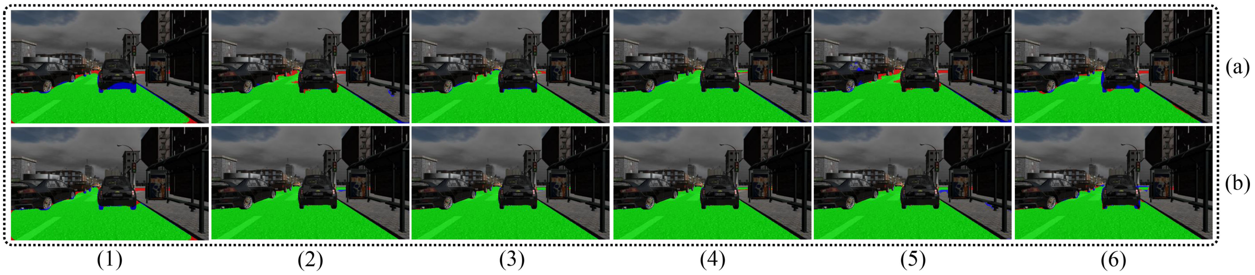

Fig. 3. Examples of the experimental results on the SYNTHIA road dataset [39], where (1) SegNet [27], (2) U-Net [28], (3) DeepLabv3+ [29], (4) DenseASPP [33], (5) DUpsampling [34], (6) GSCNN [35], (a) trained on the original training set, and (b) trained on the augmented training set. The true positive, false negative, and false positive pixels are shown in green, red, and blue, respectively.

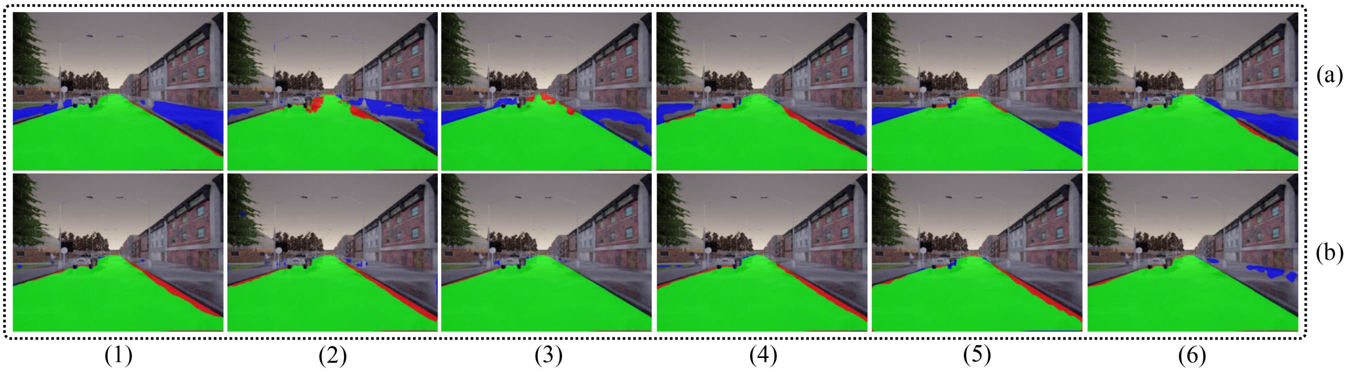

Fig. 4. Examples of the experimental results on our created SYN-Stereo road dataset, where (1) SegNet [27], (2) U-Net [28], (3) DeepLabv3+ [29], (4) DenseASPP [33], (5) DUpsampling [34], (6) GSCNN [35], (a) trained on the original training set, and (b) trained on the augmented training set. The true positive, false negative, and false positive pixels are shown in green, red, and blue, respectively.

TABLE I
PERFORMANCE COMPARISON (%) AMONG DIFFERENT DCNNS TRAINED ON THE ORIGINAL AND AUGMENTED KITTI ROAD DATASETS [17]

Network	Accuracy	Precision	Recall	F-Score	IoU
SegNet [27]	93.8	77.6	85.3	81.2	68.4
HA-SegNet	**95.6**	**85.1**	**87.3**	**86.2**	**75.7**
UNet [28]	95.7	**89.6**	82.4	85.9	75.2
HA-U-Net	**96.5**	84.4	**95.4**	**89.5**	**81.1**
DeepLabv3+ [29]	98.0	91.5	**96.4**	93.9	88.5
HA-DeepLabv3+	**98.6**	**97.2**	93.9	**95.5**	**91.4**
DenseASPP [33]	97.3	90.8	92.0	91.4	84.1
HA-DenseASPP	**98.5**	**93.9**	**96.4**	**95.1**	**90.7**
DUpsampling [34]	94.7	82.5	83.8	83.1	71.2
HA-DUpsampling	**96.2**	**90.2**	**85.2**	**87.7**	**78.0**
GSCNN [35]	94.8	84.1	82.4	83.2	71.3
HA-GSCNN	**95.4**	**87.1**	**83.2**	**85.1**	**74.1**

Best results of each network are shown in bold type.

Please note that the training, validation, and testing sets contain data from different driving scenarios, and therefore data corresponding to a single driving scenario are only contained within one of these sets.

B. Experiment Setup

In our experiments, six state-of-the-art networks: SegNet [27], U-Net [28], DeepLabv3+ [29], DenseASPP [33], DUpsampling [34], and GSCNN [35] are trained to validate the effectiveness and robustness of our proposed DS-Generator. The networks trained on the augmented training sets are named as "HA-Network," such as HA-U-Net and HA-DeepLabv3+. Furthermore, five metrics: 1) accuracy; 2) precision; 3) recall; 4) F-score; 5) the intersection over union (IoU) are used to quantify the performance of the trained DCNNs.

Additionally, other conventional training data augmentation methods, such as translation and rotation, are also used in our

TABLE II
PERFORMANCE COMPARISON (%) AMONG DIFFERENT DCNNS TRAINED ON THE ORIGINAL AND AUGMENTED SYNTHIA ROAD DATASETS [39]

Network	Accuracy	Precision	Recall	F-Score	IoU
SegNet [27]	94.1	94.5	89.5	91.9	85.1
HA-SegNet	**96.3**	**95.5**	**94.2**	**94.8**	**90.2**
UNet [28]	94.9	94.9	91.3	93.1	87.0
HA-U-Net	**97.1**	**95.8**	**96.1**	**95.9**	**92.2**
DeepLabv3+ [29]	97.2	95.0	97.4	96.2	92.7
HA-DeepLabv3+	**98.3**	**96.8**	**98.6**	**97.7**	**95.5**
DenseASPP [33]	96.0	94.0	95.1	94.5	89.7
HA-DenseASPP	**97.7**	**95.8**	**97.8**	**96.8**	**93.8**
DUpsampling [34]	95.9	95.7	93.1	94.4	89.4
HA-DUpsampling	**97.4**	**95.9**	**96.9**	**96.4**	**93.0**
GSCNN [35]	95.5	**96.4**	91.4	93.8	88.4
HA-GSCNN	**97.3**	95.3	**97.2**	**96.2**	**92.8**

Best results of each network are shown in bold type.

TABLE III
PERFORMANCE COMPARISON (%) AMONG DIFFERENT DCNNS TRAINED ON THE ORIGINAL AND AUGMENTED SYN-STEREO ROAD DATASETS

Network	Accuracy	Precision	Recall	F-Score	IoU
SegNet [27]	93.0	90.7	92.7	91.7	84.7
HA-SegNet	**95.6**	**96.6**	**93.0**	**94.8**	**90.1**
UNet [28]	92.8	90.3	92.6	91.4	84.2
HA-U-Net	**95.4**	**95.8**	**93.4**	**94.6**	**89.7**
DeepLabv3+ [29]	95.3	95.8	93.1	94.4	89.4
HA-DeepLabv3+	**97.1**	**98.2**	**95.0**	**96.6**	**93.4**
DenseASPP [33]	94.3	90.9	**95.8**	93.3	87.4
HA-DenseASPP	**96.6**	**96.8**	95.0	**95.9**	**92.1**
DUpsampling [34]	93.3	89.0	**95.3**	92.0	85.3
HA-DUpsampling	**95.9**	**96.0**	94.2	**95.1**	**90.6**
GSCNN [35]	93.8	90.8	**94.7**	92.7	86.4
HA-GSCNN	**96.4**	**97.7**	93.8	**95.7**	**91.8**

Best results of each network are shown in bold type.

TABLE IV
COMPARISONS OF THE STEREO VISION-BASED COLLISION-FREE SPACE DETECTION METHODS ON THE KITTI ROAD BENCHMARK, WHERE ↑ MEANS HIGHER VALUES ARE BETTER AND ↓ MEANS LOWER VALUES ARE BETTER

Approach	MaxF (%) ↑	AP (%) ↑	PRE (%) ↑	REC (%) ↑	FPR (%) ↓	FNR (%) ↓	Runtime (s) ↓
BM [43]	83.47	72.23	75.90	92.72	16.22	7.28	2
HistonBoost [44]	83.92	73.75	82.24	85.66	10.19	14.34	150
SCRFFPFHGSP [45]	84.93	76.31	85.37	84.49	7.98	15.51	5
GRES3D+SELAS [46]	85.09	86.86	82.27	88.10	10.46	11.90	0.11
GEO+GPR+CRF [47]	85.56	74.21	82.81	88.50	10.12	11.50	30
ProbBoost [48]	87.78	77.30	86.59	89.01	7.60	10.99	150
NNP [49]	89.68	86.50	89.67	89.68	5.69	10.32	5
BMCF [50]	89.75	84.15	89.02	90.49	6.15	9.51	2.50
HA-DeepLabv3+ (Ours)	**94.83**	**93.24**	**94.77**	**94.89**	**2.88**	**5.11**	**0.06**

Best results are shown in bold type.

experiments. The stochastic gradient descent with momentum optimizer is utilized to minimize the loss function, and the initial learning rate is set to 0.001. Furthermore, we adopt the early-stopping mechanism [42] on the validation set to reduce over-fitting problem. The DCNN performance is then quantified on the testing set, as presented in Section IV-C. Moreover, we select the best-performing model and fine-tune it for the result submission to the KITTI road benchmark [17].

C. Performance Evaluation

This subsection evaluates the performance of our proposed DS-Generator both qualitatively and quantitatively. Examples of the experimental results on the KITTI [17], SYNTHIA [39], and our SYN-Stereo road datasets are shown in Figs. 2–4, respectively. We can clearly observe that the DCNNs trained on the augmented training set generally perform better

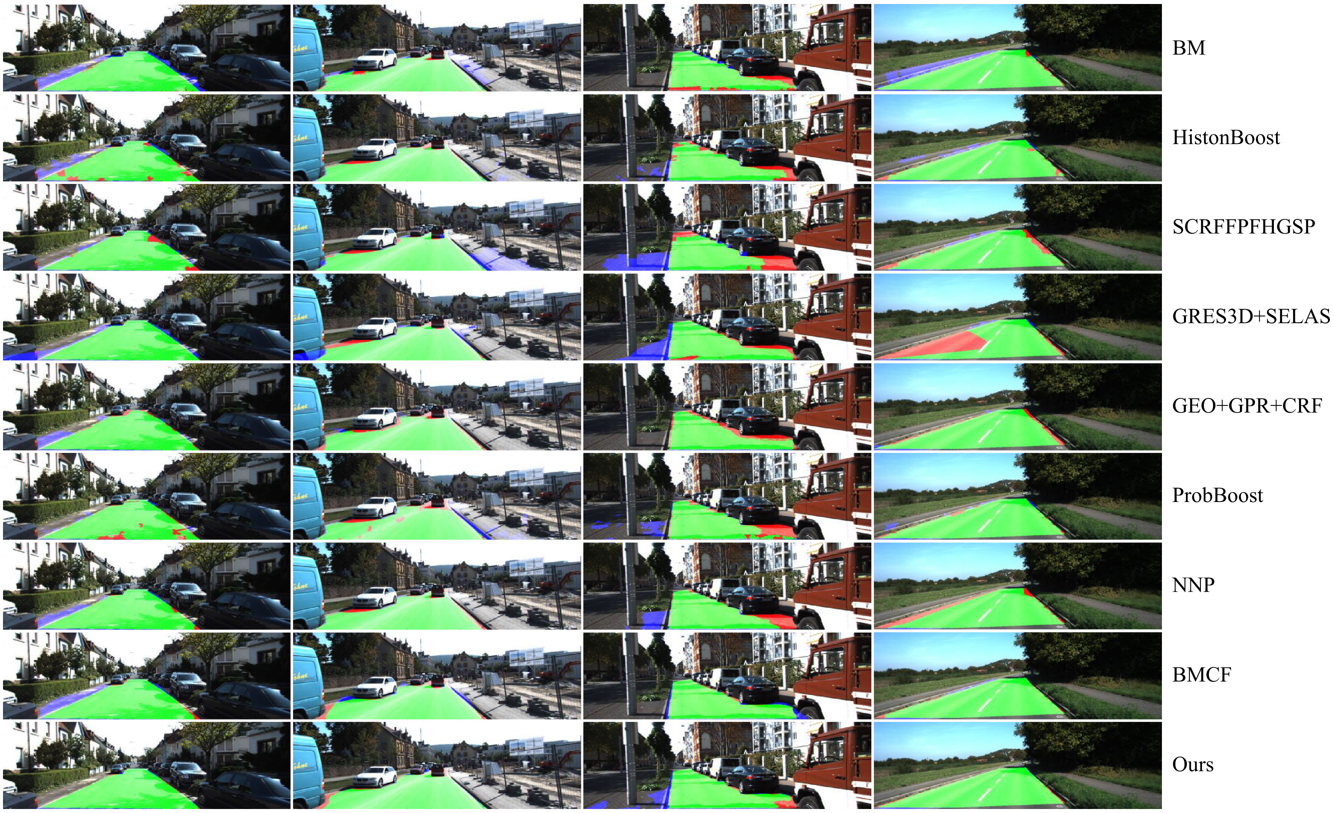

Fig. 5. Examples of the experimental results on the KITTI road benchmark, where the true positive, false negative, and false positive pixels are shown in green, red, and blue, respectively.

than the same DCNNs trained on the original training set. The corresponding quantitative comparisons are given in Tables I–III, respectively, where it can be seen that the F-score and IoU of the DCNNs trained on the augmented training set obtained by our proposed DS-Generator are improved by around 1.5%–5.0% and 2.8%–7.3%, respectively. Furthermore, HA-DeepLabv3+ performs better than all other DCNNs. Our analysis shows that, compared to the common training set augmentation operations, our proposed DS-Generator can leverage the relationship between multiview images to perform more effective training data augmentation, and thus, benefit all state-of-the-art DCNNs for collision-free space detection.

As mentioned above, we fine-tune our best-performing method, HA-DeepLabv3+,[4] and submit its results to the KITTI road benchmark [17]. Then, we compare our HA-DeepLabv3+ with eight state-of-the-art stereo vision-based collision-free space detection methods: BM [43], HistonBoost [44], SCRFF-PFHGSP [45], GRES3D+SELAS [46], GEO+GPR+CRF [47], ProbBoost [48], NNP [49], and BMCF [50], published on the KITTI road benchmark. Examples of the experimental results are shown in Fig. 5. The quantitative comparisons are given in Table IV. Readers can see that our HA-DeepLabv3+ is the best stereo vision-based collision-free space detection method, which achieves the highest MaxF (maximum F-score), AP (average precision), PRE (precision), REC (recall), FPR (false positive rate) and FNR (false negative rate). Furthermore, our method runs in real time and it is much faster than all other compared methods.

V. Conclusion

This article proposed a novel training data augmentation approach, referred to as DS-Generator. It can generate additional driving scene images from multiview vision data, such as stereo image pairs. Furthermore, we published a synthetic collision-free space detection dataset, named SYN-Stereo road dataset for research purposes. Extensive experimental results conducted with six state-of-the-art DCNNs on three datasets demonstrated the effectiveness of our DS-Generator, where the F-score and IoU of the DCNNs are improved by around 1.5%–5.0% and 2.8%–7.3%, respectively. Furthermore, HA-DeepLabv3+, our best-performing implementation, achieves the best overall performance compared to other stereo vision-based collision-free space detection algorithms published on the KITTI road benchmark.

References

[1] F. Pieri, C. Zambelli, A. Nannini, P. Olivo, and S. Saponara, “Is consumer electronics redesigning our cars?: Challenges of integrated technologies for sensing, computing, and storage,” *IEEE Consum. Electron. Mag.*, vol. 7, no. 5, pp. 8–17, Sep. 2018.

[4] [Online]. Available: www.cvlibs.net/datasets/kitti/eval_road_detail.php?result=4d39ae0a09df67b61c037ad3829f1a2c2b848f07

[2] J. Zhang and K. B. Letaief, "Mobile edge intelligence and computing for the internet of vehicles," *Proc. IEEE*, vol. 108, no. 2, pp. 246–261, Feb. 2020.

[3] R. Fan, L. Wang, M. J. Bocus, and I. Pitas, "Computer stereo vision for autonomous driving," *CoRR*, 2020.

[4] E. Stewart, "Self-driving cars have to be safer than regular cars. The question is how much," *Vox*, vol. 17, May 2019.

[5] M. Nagai, "Research into ADAS with autonomous driving intelligence for future innovation," in *Proc. 5th Int. Munich Chassis Symp.*, 2014, pp. 779–793.

[6] W. Biever, L. Angell, and S. Seaman, "Automated driving system collisions: Early lessons," *Human Factors*, vol. 62, no. 2, pp. 249–259, 2020.

[7] S. Grigorescu, B. Trasnea, T. Cocias, and G. Macesanu, "A survey of deep learning techniques for autonomous driving," *J. Field Robot.*, vol. 37, no. 3, pp. 362–386, 2020.

[8] L. Sless, B. El Shlomo, G. Cohen, and S. Oron, "Road scene understanding by occupancy grid learning from sparse radar clusters using semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, 2019, pp. 867–875.

[9] H. Wang, R. Fan, Y. Sun, and M. Liu, "Dynamic fusion module evolves drivable area and road anomaly detection: A benchmark and algorithms," *IEEE Trans. Cybern.*, 2021, doi: 10.1109/TCYB.2021.3064089.

[10] L. A. Thiede and P. P. Brahma, "Analyzing the variety loss in the context of probabilistic trajectory prediction," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 9954–9963.

[11] R. Fan and N. Dahnoun, "Real-time stereo vision-based lane detection system," *Meas. Sci. Technol.*, vol. 29, no. 7, 2018, Art. no. 074005.

[12] S. Pouyanfar, M. Saleem, N. George, and S.-C. Chen, "Roads: Randomization for obstacle avoidance and driving in simulation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2019, pp. 1267–1276.

[13] 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.

[14] H. Wang, R. Fan, Y. Sun, and M. Liu, "Applying surface normal information in drivable area and road anomaly detection for ground mobile robots," *Int. Conf. Intell. Robot. Syst.*, 2020, doi: 10.1109/IROS45743.2020.9341340.

[15] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*. Cambridge, U.K.: Cambridge Univ. Press, 2003.

[16] R. Fan, X. Ai, and N. Dahnoun, "Road surface 3 D reconstruction based on dense subpixel disparity map estimation," *IEEE Trans. Image Process.*, vol. 27, no. 6, pp. 3025–3035, Jun. 2018.

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

[18] R. Fan, J. Jiao, J. Pan, H. Huang, S. Shen, and M. Liu, "Real-time dense stereo embedded in a UAV for road inspection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2019, pp. 535–543.

[19] 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, no. 1, pp. 897–908, Aug. 2019.

[20] A. Wedel, H. Badino, C. Rabe, H. Loose, U. Franke, and D. Cremers, "B-spline modeling of road surfaces with an application to free-space estimation," *IEEE Trans. Intell. Transp. Syst.*, vol. 10, no. 4, pp. 572–583, Dec. 2009.

[21] G. D. Knott, *Interpolating Cubic Splines*. New York, NY, USA: Springer-Verlag, vol. 18, 2000.

[22] R. Labayrade, D. Aubert, and J.-P. Tarel, "Real time obstacle detection in stereovision on non flat road geometry through "V-disparity" representation," in *Proc. Intell. Veh. Symp.*, 2002, vol. 2, 646–651.

[23] Y. Zhang, Y. Su, J. Yang, J. Ponce, and H. Kong, "When Dijkstra meets vanishing point: A stereo vision approach for road detection," *IEEE Trans. Image Process.*, vol. 27, no. 5, pp. 2176–2188, May 2018.

[24] A. Goldberg and T. Radzik, "A heuristic improvement of the Bellman-Ford algorithm," *Appl. Math. Lett.*, vol. 6, no. 3, pp. 3–6, May 1993.

[25] 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.

[26] 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.

[27] 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.

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

[29] 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.

[30] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected CRFs," in *Proc. Int. Conf. Learn. Representations*, 2015.

[31] Chen, G. Liang-Chieh, I. Papandreou, K. Kokkinos Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2017.

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

[33] 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.

[34] 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.

[35] 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. 5229–5238.

[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. Workshop*, 2020, pp. 285–300.

[37] 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.

[38] 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," 2021, *arXiv:2012.10802*.

[39] D. Hernandez-Juarez *et al.*, "Slanted stixels: Representing San Francisco's steepest streets," in *Proc. Brit. Mach. Vis. Conf.*, 2017, pp. 87.1–87.12.

[40] J.-R. Chang and Y.-S. Chen, "Pyramid stereo matching network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 5410–5418.

[41] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "CARLA: An open urban driving simulator," in *Proc. 1st Annu. Conf. Robot Learn.,* Nov. 2017, pp. 1–16.

[42] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep Learning*. Cambridge, MA, USA: MIT Press, vol. 1, no. 2, 2016.

[43] B. Wang, V. Frémont, and S. A. Rodríguez, "Color-based road detection and its evaluation on the kitti road benchmark," in *Proc. IEEE Intell. Veh. Symp.*, 2014, pp. 31–36.

[44] G. B. Vitor, A. C. Victorino, and J. V. Ferreira, "Comprehensive performance analysis of road detection algorithms using the common urban kitti-road benchmark," in *Proc. IEEE Intell. Veh. Symp.*, 2014, pp. 19–24.

[45] I. V. Gheorghe, "Semantic segmentation of terrain and road terrain for advanced driver assistance systems," Ph.D. dissertation, Comput. Sci. Autom. Eng., Coventry Univ., Coventry, U.K., 2015.

[46] P. Y. Shinzato, "Estimation of obstacles and road area with sparse 3D points," Inst. Math. Comput. Sci., Univ. Sao Paulo, Sao Paulo, Brazil, 2015.

[47] Z. Xiao *et al.*, "Gaussian process regression-based robust free space detection for autonomous vehicle by 3-D point cloud and 2-D appearance information fusion," *Int. J. Adv. Robot. Syst.*, vol. 14, no. 4, 2017, Art. no. 1729881417717058.

[48] G. B. Vitor, A. C. Victorino, and J. V. Ferreira, "A probabilistic distribution approach for the classification of urban roads in complex environments," in *Proc. IEEE Proc. ICRA Workshop Workshop Model., Estimation, Perception Control All Terrain Mobile Robots*, 2014.

[49] X. Chen *et al.*, "3D object proposals for accurate object class detection," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 424–432.

[50] L. Wang, T. Wu, Z. Xiao, L. Xiao, D. Zhao, and J. Han, "Multi-cue road boundary detection using stereo vision," in *Proc. IEEE Int. Conf. Veh. Electron. Saf.*, 2016, pp. 1–6.

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

Between July 2018 and February 2020, he was the Deputy Director of Robotics and Multiperception Laboratory (RAM-Lab), as well as a Research Associate with the Robotics Institute and the Department of Electronic and Computer Engineering, Hong Kong University of Science and Technology, Hong Kong. He co-found ATG Robotics in July 2019 and has been working as their Chief Scientist since September 2019. Since February 2020, he has been a Postdoc Fellow with the Department of Ophthalmology and the Department of Computer Science and Engineering, University of California San Diego, La Jolla, CA, USA. He will join the Department of Control Science and Engineering, the School of Electronic and Information Engineering, and Shanghai Research Institute for Intelligent Autonomous Systems, Tongji University, Shanghai, China, as a Research Full Professor in Fall 2021. He is the Director of Machine Intelligence and Autonomous Systems (MIAS) research group. His research interests include computer vision, machine/deep learning, image/signal processing, autonomous driving, and bio-informatics.

Dr. Fan is the General Chair of Autonomous Vehicle Vision (AVVision) community (www.avvision.xyz).

Hengli Wang (Graduate Student Member, IEEE) received the B.E. degree in mechatronics engineering from Zhejiang University, Hangzhou, China, in 2018. He is currently working toward the Ph.D. degree in electronic and computer engineering with the Robotics Institute, The Hong Kong University of Science and Technology (HKUST), Hong Kong.

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

Peide Cai received the B.E. degree in automation from Zhejiang University, Hangzhou, China, in 2018. He is currently working toward the Ph.D. degree in electronic and computer engineering with the Robotics Institute, The Hong Kong University of Science and Technology (HKUST), Hong Kong.

His research interests include robot navigation, machine learning, and sensor fusion.

Jin Wu (Member, IEEE) is currently working toward the Ph.D. degree in electronic and computer engineering with the Robotics Institute, Hong Kong University of Science and Technology (HKUST), Hong Kong.

He has been a Research Assistant with the Department of Electronic and Computer Engineering, HKUST. His research interests include robot navigation, multisensor fusion, automatic control, and mechatronics.

Mohammud Junaid Bocus received the B.Eng. degree (with first-class honors) in electronic and communication engineering from the University of Mauritius, Moka, Mauritius, in 2012, and the M.Sc. (distinction) degree in wireless communications and signal processing and the Ph.D. degree in electrical and electronic engineering from the University of Bristol, Bristol, U.K, in 2015 and 2020, respectively.

He is currently working as a Research Associate with the University of Bristol, focusing on activity recognition and localisation using commodity WiFi and ultrawideband (UWB) systems. His research interests include computer stereo vision, terrestrial and underwater wireless communication based on orthogonal frequency-division multiplexing (OFDM), filter-bank multicarrier (FBMC) modulation, multiple-input–multiple-output (MIMO), and massive MIMO systems as well as video coding and machine/deep learning.

Lei Qiao received the B.S. degree in automation and the M.Eng. degree in control engineering from the College of Automation, Harbin Engineering University, Harbin, China, in 2012 and 2014, respectively, and the Ph.D. degree in control science and engineering from the Department of Automation, Shanghai Jiao Tong University, Shanghai, China, in 2020.

From September 2017 to September 2018, he was a Visiting Research Scholar with the Department of Electrical and Computer Engineering, Ohio State University, Columbus, OH, USA. Currently, he is an Assistant Professor with the School of Naval Architecture, Ocean and Civil Engineering, as well as the State Key Laboratory of Ocean Engineering, Shanghai Jiao Tong University. His research interests include navigation and control of autonomous robotics such as autonomous underwater vehicles, unmanned surface vehicles and unmanned aerial vehicles, and multiple autonomous robotics coordination including across heterogeneous domains coordination.

Dr. Qiao won the 2019 Premium Award for Best Paper in IET Control Theory and Applications and the 2020 Best Paper Award of Shanghai Association of Automation, both as the rst author. He is the outstanding Ph.D. graduates of Shanghai, in 2020.

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

He is involved in several NSF projects, and National 863-Hi-Tech-Plan projects in China. He is the PI of over 20 projects including projects funded by RGC, NSFC, ITC, SZSTI, etc. He has authored or coauthored over 90 papers in major international journals and conferences. His research interests include dynamic environment modeling, 3-D mapping, machine learning, and visual control.

Dr. Liu won the second place of EMAV 2009 (European Micro Aerial Vehicle Competition). He got two awards from IARC 14 (International Aerial Robotics Contest). He won the Best Student Paper Award as the first author for MFI 2012 (IEEE International Conference on Multisensor Fusion and Information Integration), the Best Paper Award in Information for ICIA 2013 (IEEE International Conference on Information and Automation) as first author and Best Paper Award Finalists as co-author, the Best RoboCup Baper Award for IROS 2013 (IEEE/RSJ International Conference on Intelligent Robots and Systems). He won the Best Student Paper Award of IEEE ICAR 2017, the Best Paper Award in Automation for ICIA 2017. He won twice the innovation contest Chunhui Cup Winning Award, in 2012 and 2013. He won the Wu Wenjun AI Innovation Award in 2016. He was the General Chair of International Conference on Computer Vision Systems (ICVS) 2017; the Program Chair of IEEE International Conference on Real-time Computing and Robotics (IEEE-RCAR) 2016; the Program Chair of International Robotic Alliance Conference 2017. He is the Awardee of IEEE IROS Toshio Fukuda Young Professional Award 2018.