SNE-RoadSeg: Incorporating Surface Normal Information into Semantic Segmentation for Accurate Freespace Detection

Rui Fan[1⋆][0000–0003–2593–6596], Hengli Wang[2⋆][0000–0002–7515–9759], Peide Cai[2][0000–0002–9759–2991], and Ming Liu[2][0000–0002–4500–238X]

[1] UC San Diego
rui.fan@ieee.org
[2] HKUST Robotics Institute
{hwangdf, peide.cai, eelium}@ust.hk

Abstract. Freespace detection is an essential component of visual perception for self-driving cars. The recent efforts made in data-fusion convolutional neural networks (CNNs) have significantly improved semantic driving scene segmentation. Freespace can be hypothesized as a ground plane, on which the points have similar surface normals. Hence, in this paper, we first introduce a novel module, named surface normal estimator (SNE), which can infer surface normal information from dense depth/disparity images with high accuracy and efficiency. Furthermore, we propose a data-fusion CNN architecture, referred to as RoadSeg, which can extract and fuse features from both RGB images and the inferred surface normal information for accurate freespace detection. For research purposes, we publish a large-scale synthetic freespace detection dataset, named Ready-to-Drive (R2D) road dataset, collected under different illumination and weather conditions. The experimental results demonstrate that our proposed SNE module can benefit all the state-of-the-art CNNs for freespace detection, and our SNE-RoadSeg achieves the best overall performance among different datasets.

Keywords: freespace detection · self-driving cars · data-fusion CNN · semantic driving scene segmentation · surface normal

Source Code, Dataset and Demo Video:
sites.google.com/view/sne-roadseg

1 Introduction

Autonomous cars are a regular feature in science fiction films and series, but thanks to the rise of artificial intelligence, the fantasy of picking up one such vehicle at your garage forecourt has turned into reality. Driving scene understanding is a crucial task for autonomous cars, and it has taken a big leap

[⋆] These authors contributed equally to this work and are therefore joint first authors.

with recent advances in artificial intelligence [1]. Collision-free space (or simply freespace) detection is a fundamental component of driving scene understanding [27]. Freespace detection approaches generally classify each pixel in an RGB or depth/disparity image as drivable or undrivable. Such pixel-level classification results are then utilized by other modules in the autonomous system, such as trajectory prediction [4] and path planning [31], to ensure that the autonomous car can navigate safely in complex environments.

The existing freespace detection approaches can be categorized as either traditional or machine/deep learning-based. The traditional approaches generally formulate freespace with an explicit geometry model and find its best coefficients using optimization approaches [13]. [36] is a typical traditional freespace detection algorithm, where road segmentation is performed by fitting a B-spline model to the road disparity projections on a 2D disparity histogram (generally known as a v-disparity image) [12]. With recent advances in machine/deep learning, freespace detection is typically regarded as a semantic driving scene segmentation problem, where the convolutional neural networks (CNNs) are used to learn its best solution [34]. For instance, Lu *et al.* [25] employed an encoder-decoder architecture to segment RGB images in the bird's eye view for end-to-end freespace detection. Recently, many researchers have resorted to data-fusion CNN architectures to further improve the accuracy of semantic image segmentation. For example, Hazirbas *et al.* [19] incorporated depth information into conventional semantic segmentation via a data-fusion CNN architecture, which greatly enhanced the performance of driving scene segmentation.

In this paper, we first introduce a novel module named surface normal estimator (SNE), which can infer surface normal information from dense disparity/depth images with both high precision and efficiency. Additionally, we design a data-fusion CNN architecture named RoadSeg, which is capable of incorporating both RGB and surface normal information into semantic segmentation for accurate freespace detection. Since the existing freespace detection datasets with diverse illumination and weather conditions do not have either disparity/depth information or freespace ground truth, we created a large-scale synthetic freespace detection dataset, named Ready-to-Drive (R2D) road dataset (containing 11430 pairs of RGB and depth images), under different illumination and weather conditions. Our R2D road dataset is also publicly available for research purposes. To validate the feasibility and effectiveness of our introduced SNE module, we use three road datasets (KITTI [15], SYNTHIA [21] and our R2D) to train ten state-of-the-art CNNs (six single-modal CNNs and four data-fusion CNNs), with and without our proposed SNE module embedded. The experiments demonstrate that our proposed SNE module can benefit all these CNNs for freespace detection. Also, our method SNE-RoadSeg outperforms all other CNNs for freespace detection, where its overall performance is the second best on the KITTI road benchmark[3] [15].

The remainder of this paper is structured as follows: Section 2 provides an overview of the state-of-the-art CNNs for semantic image segmentation. Section

[3] cvlibs.net/datasets/kitti/eval_road.php

3 introduces our proposed SNE-RoadSeg. Section 4 shows the experimental results and discusses both the effectiveness of our proposed SNE module and the performance of our SNE-RoadSeg. Finally, Section 5 concludes the paper.

2 Related Work

In 2015, Long *et al.* [24] introduced Fully Convolutional Network (FCN), a CNN for end-to-end semantic image segmentation. Since then, research on this topic has exploded. Based on FCN, Ronneberger *et al.* [26] proposed U-Net in the same year, which consists of a contracting path and an expansive path [26]. It adds skip connections between the contracting path and the expansive path to help better recover the full spatial resolution. Different from FCN, SegNet [3] utilizes an encoder-decoder architecture, which has become the mainstream structure for following approaches. An encoder-decoder architecture is typically composed of an encoder, a decoder and a final pixel-wise classification layer.

Furthermore, DeepLabv3+ [9], developed from DeepLabv1 [6], DeepLabv2 [7] and DeepLabv3 [8], was proposed in 2018. It employs depth-wise separable convolution in both atrous spatial pyramid pooling (ASPP) and the decoder, which makes its encoder-decoder architecture much faster and stronger [9]. Although the ASPP can generate feature maps by concatenating multiple atrous-convolved features, the resolution of the generated feature maps is not sufficiently dense for some applications such as autonomous driving [7]. To address this problem, DenseASPP [37] was designed to connect atrous convolutional layers (ACLs) densely. It is capable of generating multi-scale features that cover a larger and denser scale range, without significantly increasing the model size [37].

Different from the above-mentioned CNNs, DUpsampling [32] was proposed to recover the pixel-wise prediction by employing a data-dependent decoder. It allows the decoder to downsample the fused features before merging them, which not only reduces computational costs, but also decouples the resolutions of both the fused features and the final prediction [32]. GSCNN [30] utilizes a novel two-branch architecture consisting of a regular (classical) branch and a shape branch. The regular branch can be any backbone architecture, while the shape branch processes the shape information in parallel with the regular branch. Experimental results have demonstrated that this architecture can significantly boost the performance on thinner and smaller objects [30].

FuseNet [19] was designed to use RGB-D data for semantic image segmentation. The key ingredient of FuseNet is a fusion block, which employs element-wise summation to combine the feature maps obtained from two encoders. Although FuseNet [19] demonstrates impressive performance, the ability of CNNs to handle geometric information is limited, due to the fixed grid kernel structure [35]. To address this problem, depth-aware CNN [35] presents two intuitive and flexible operations: depth-aware convolution and depth-aware average pooling. These operations can efficiently incorporate geometric information into the CNN by leveraging the depth similarity between pixels [35].

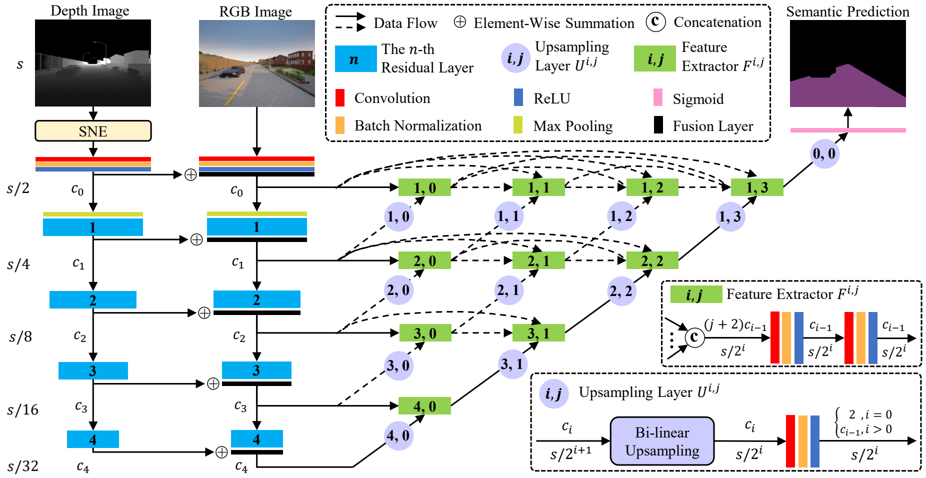


Fig. 1: The architecture of our SNE-RoadSeg. It consists of our SNE module, an RGB encoder, a surface normal encoder and a decoder with densely-connected skip connections. s represents the input resolution of the RGB and depth images. c_n represents the number of feature map channels at different levels.

MFNet [18] was proposed for semantic driving scene segmentation with the use of RGB-thermal vision data. In order to meet the real-time requirement of autonomous driving applications, MFNet focuses on minimizing the trade-off between accuracy and efficiency. Similarly, RTFNet [29] was developed to improve the semantic image segmentation performance using RGB-thermal vision data. Its main contribution is a novel decoder, which leverages short-cuts to produce sharp boundaries while keeping more detailed information [29].

3 SNE-RoadSeg

3.1 SNE

The proposed SNE is developed from our recent work three-filters-to-normal (3F2N) [14]. Its architecture is shown in Fig. 2. For a perspective camera model, a 3D point $\mathbf{P} = [X, Y, Z]^\top$ in the Euclidean coordinate system can be linked with a 2D image pixel $\mathbf{p} = [x, y]^\top$ using:

$$Z \begin{bmatrix} \mathbf{p} \\ 1 \end{bmatrix} = \mathbf{K}\mathbf{P} = \begin{bmatrix} f_x & 0 & x_o \\ 0 & f_y & y_o \\ 0 & 0 & 1 \end{bmatrix} \mathbf{P}, \tag{1}$$

where $\mathbf{K}$ is the camera intrinsic matrix; $\mathbf{p}_o = [x_o, y_o]^\top$ is the image center; f_x and f_y are the camera focal lengths in pixels. The simplest way to estimate the surface normal $\mathbf{n} = [n_x, n_y, n_z]^\top$ of $\mathbf{P}$ is to fit a local plane:

$$n_x X + n_y Y + n_z Z + d = 0 \tag{2}$$

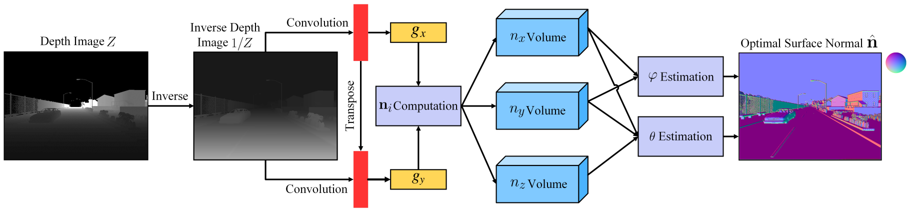


Fig. 2: The architecture of our proposed SNE module.

to $\mathbf{N}_{\mathbf{P}}^{+} = [\mathbf{P}, \mathbf{N}_{\mathbf{P}}]^\top$, where $\mathbf{N}_{\mathbf{P}} = [\mathbf{Q}_1, \dots, \mathbf{Q}_k]^\top$ is a set of k neighboring points of $\mathbf{P}$. Combining (1) and (2) results in [14]:

$$\frac{1}{Z} = -\frac{1}{d}\left(n_x \frac{x - x_o}{f_x} + n_y \frac{y - y_o}{f_y} + n_z\right). \tag{3}$$

Differentiating (3) with respect to x and y leads to:

$$g_x = \frac{\partial 1/Z}{\partial x} = -\frac{n_x}{d f_x}, \quad g_y = \frac{\partial 1/Z}{\partial y} = -\frac{n_y}{d f_y}, \tag{4}$$

which, as illustrated in Fig. 2, can be respectively approximated by convolving the inverse depth image $1/Z$ (or a disparity image, as disparity is in inverse proportion to depth) with a horizontal and a vertical image gradient filter [14]. Rearranging (4) results in the expressions of n_x and n_y as follows:

$$n_x = -d f_x g_x, \quad n_y = -d f_y g_y. \tag{5}$$

Given an arbitrary $\mathbf{Q}_i \in \mathbf{N}_{\mathbf{P}}$, we can compute its corresponding n_{z_i} by plugging (5) into (2):

$$n_{z_i} = d\frac{f_x \Delta X_i g_x + f_y \Delta Y_i g_y}{\Delta Z_i}, \tag{6}$$

where $\mathbf{Q}_i - \mathbf{P} = [\Delta X_i, \Delta Y_i, \Delta Z_i]^\top$. Since (5) and (6) have a common factor of $-d$, the surface normal $\mathbf{n}_i$ obtained from $\mathbf{Q}_i$ and $\mathbf{P}$ has the following expression [34]:

$$\mathbf{n}_i = \left[f_x g_x, \quad f_y g_y, \quad -\frac{f_x \Delta X_i g_x + f_y \Delta Y_i g_y}{\Delta Z_i}\right]^\top. \tag{7}$$

A k-connected neighborhood system $\mathbf{N}_{\mathbf{P}}$ of $\mathbf{P}$ can produce k normalized surface normals $\bar{\mathbf{n}}_1, \dots, \bar{\mathbf{n}}_k$, where $\bar{\mathbf{n}}_i = \frac{\mathbf{n}_i}{\|\mathbf{n}_i\|_2} = [\bar{n}_{x_i}, \bar{n}_{y_i}, \bar{n}_{z_i}]^\top$. Since any normalized surface normals are projected on a sphere with center $(0, 0, 0)$ and radius 1, we believe that the optimal surface normal $\hat{\mathbf{n}}$ for $\mathbf{P}$ is also projected somewhere on the same sphere, where the projections of $\bar{\mathbf{n}}_1, \dots, \bar{\mathbf{n}}_k$ distribute most intensively [13]. $\hat{\mathbf{n}}$ can be written in spherical coordinates as follows:

$$\hat{\mathbf{n}} = \left[\sin\theta\cos\varphi, \ \sin\theta\sin\varphi, \ \cos\theta\right]^\top, \tag{8}$$

where $\theta \in [0, \pi]$ denotes inclination and $\varphi \in [0, 2\pi)$ denotes azimuth. φ can be computed using:

$$\varphi = \arctan\left(\frac{f_y g_y}{f_x g_x}\right). \tag{9}$$

Similar to [13], we hypothesize that the angle between an arbitrary pair of normalized surface normals is less than $\pi/2$. $\hat{\mathbf{n}}$ can therefore be estimated by minimizing $E = -\sum_{i=1}^{k} \hat{\mathbf{n}} \cdot \bar{\mathbf{n}}_i$ [13]. $\frac{\partial E}{\partial \theta} = 0$ obtains:

$$\theta = \arctan\left(\frac{\sum_{i=1}^{k} \bar{n}_{x_i} \cos\varphi + \sum_{i=1}^{k} \bar{n}_{y_i} \sin\varphi}{\sum_{i=1}^{k} \bar{n}_{z_i}}\right). \tag{10}$$

Substituting θ and φ into (8) results in the optimal surface normal $\hat{\mathbf{n}}$, as shown in Fig. 2. The performance of our proposed SNE will be discussed in Section 4.

3.2 RoadSeg

U-Net [26] has demonstrated the effectiveness of using skip connections in recovering the full spatial resolution. However, its skip connections force aggregations only at the same-scale feature maps of the encoder and decoder, which, we believe, is an unnecessary constraint. Inspired by DenseNet [23], we propose RoadSeg, which exploits densely-connected skip connections to realize flexible feature fusion in the decoder.

As shown in Fig. 1, our proposed RoadSeg also adopts the popular encoder-decoder architecture. An RGB encoder and a surface normal encoder is employed to extract the feature maps from RGB images and from the inferred surface normal information, respectively. The extracted RGB and surface normal feature maps are hierarchically fused through element-wise summations. The fused feature maps are then fused again in the decoder through densely-connected skip connections to restore the resolution of the feature maps. At the end of RoadSeg, a sigmoid layer is used to generate the probability map for the semantic driving scene segmentation.

We use ResNet [20] as the backbone of our RGB and surface normal encoders, the structures of which are identical to each other. Specifically, the initial block consists of a convolutional layer, a batch normalization layer and a ReLU activation layer. Then, a max pooling layer and four residual layers are sequentially employed to gradually reduce the resolution as well as increase the number of feature map channels. ResNet has five architectures: ResNet-18, ResNet-34, ResNet-50, ResNet-101 and ResNet-152. Our RoadSeg follows the same naming rule of ResNet. c_n, the number of feature map channels (see Fig. 1) varies with respect to the adopted ResNet architecture. Specifically, c_0–c_4 are 64, 64, 128, 256 and 512, respectively, for ResNet-18 and ResNet-34, and are 64, 256, 512, 1024 and 2048, respectively, for ResNet-50, ResNet-101 and ResNet-152.

The decoder consists of two different types of modules: a) feature extractors $F^{i,j}$ and b) upsampling layers $U^{i,j}$, which are connected densely to realize flexible feature fusion. The feature extractor is employed to extract features from the fused feature maps, and it ensures that the feature map resolution is unchanged. The upsampling layer is employed to increase the resolution and decrease the feature map channels. Three convolutional layers in the feature extractor and the upsampling layer have the same kernel size of 3×3, the same stride of 1 and the same padding of 1.

4 Experiments

4.1 Datasets and Experimental Setup

In our experiments, we first evaluate the performance of our proposed SNE on the DIODE dataset [33], a public surface normal estimation dataset containing RGBD vision data of both indoor and outdoor scenarios. We utilize the average angular error (AAE), $e_{\mathrm{AAE}} = \frac{1}{m}\sum_{k=1}^{m}\cos^{-1}\left(\frac{\langle \mathbf{n}_k, \hat{\mathbf{n}}_k \rangle}{\|\mathbf{n}_k\|_2 \|\hat{\mathbf{n}}_k\|_2}\right)$, to quantify our SNE's accuracy, where m is the number of 3D points used for evaluation; $\mathbf{n}_k$ and $\hat{\mathbf{n}}_k$ is the ground truth and estimated (optimal) surface normal, respectively. The experimental results are presented in Section 4.2.

Then, we carry out the experiments on the following three datasets to evaluate the performance of our proposed SNE-RoadSeg for freespace detection:

- The KITTI road dataset [15]: this dataset provides real-world RGB-D vision data. We split it into three subsets: a) training (173 images), b) validation (58 images), and c) testing (58 images).
- The SYNTHIA road dataset [21]: this dataset provides synthetic RGB-D vision data. We select 2224 images from it and group them into: a) training (1334 images), b) validation (445 images), and c) testing (445 images).
- Our R2D road dataset: along with our proposed SNE-RoadSeg, we also publish a large-scale synthetic freespace detection dataset, named R2D road dataset. This dataset is created using the CARLA[4] simulator [11]. Firstly, we mount a simulated stereo rig (baseline: 1.5 m) on the top of a vehicle to capture synchronized stereo images (resolution: 640×480 pixels) at 10 fps. The vehicle navigates in six different scenarios under different illumination and weather conditions (sunny, rainy, day and sunset). There are a total of 11430 pairs of stereo images with corresponding depth images and semantic segmentation ground truth. We split them into three subsets: a) training (6117 images), b) validation (2624 images), and c) testing (2689 images). Our dataset is publicly available at `sites.google.com/view/sne-roadseg` for research purposes.

We use these three datasets to train ten state-of-the-art CNNs, including six single-modal CNNs and four data-fusion CNNs. We conduct the experiments of single-modal CNNs with three setups: a) training with RGB images, b) training with depth images, and c) training with surface normal images (generated from depth images using our SNE), which are denoted as **RGB**, **Depth** and

[4] `carla.org`

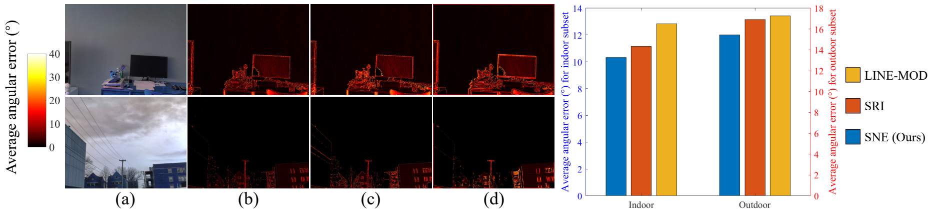


Fig. 3: Qualitative and quantitative results on the DIODE dataset: (a) RGB images; (b)–(d): the angular error maps obtained using our proposed SNE, SRI [2] and LINE-MOD [22], respectively.

SNE-Depth, respectively. Similarly, the experiments of data-fusion CNNs are conducted using two setups: training using RGB-D vision data, with and without our SNE embedded, which are denoted as **RGBD** and **SNE-RGBD**, respectively. To compare the performances between our proposed RoadSeg and other state-of-the-art CNNs, we train our RoadSeg with the same setups as for the data-fusion CNNs on the three datasets. Moreover, we re-train our SNE-RoadSeg for the result submission to the KITTI road benchmark [15]. The experimental results are presented in Section 4.3. Additionally, the ablation study of our SNE-RoadSeg is provided in Section 4.4.

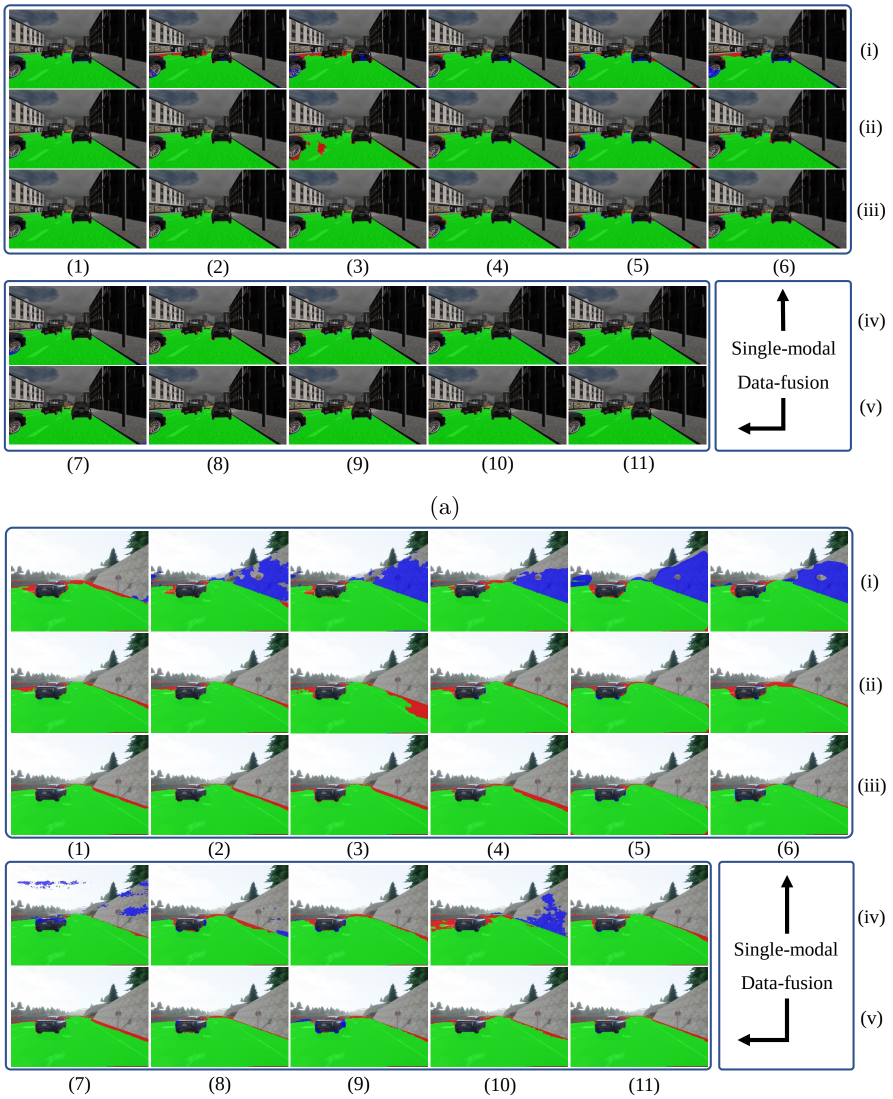
Five common metrics are used for the performance evaluation of freespace detection: accuracy, precision, recall, F-score and the intersection over union (IoU). Their corresponding definitions are as follows: Accuracy $= \frac{n_{\mathrm{tp}}+n_{\mathrm{tn}}}{n_{\mathrm{tp}}+n_{\mathrm{tn}}+n_{\mathrm{fp}}+n_{\mathrm{fn}}}$, Precision $= \frac{n_{\mathrm{tp}}}{n_{\mathrm{tp}}+n_{\mathrm{fp}}}$, Recall $= \frac{n_{\mathrm{tp}}}{n_{\mathrm{tp}}+n_{\mathrm{fn}}}$, F-score $= \frac{2n_{\mathrm{tp}}^2}{2n_{\mathrm{tp}}^2+n_{\mathrm{tp}}\left(n_{\mathrm{fp}}+n_{\mathrm{fn}}\right)}$ and IoU $= \frac{n_{\mathrm{tp}}}{n_{\mathrm{tp}}+n_{\mathrm{fp}}+n_{\mathrm{fn}}}$, where n_{tp}, n_{tn}, n_{fp} and n_{fn} represents the true positive, true negative, false positive, and false negative pixel numbers, respectively. In addition, the stochastic gradient descent with momentum (SGDM) 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 on the validation subset to avoid over-fitting. The performance is then quantified using the testing subset.

4.2 Performance Evaluation of Our SNE

We simply set $g_x = \frac{1}{Z(x-1,y)} - \frac{1}{Z(x+1,y)}$ and $g_y = \frac{1}{Z(x,y-1)} - \frac{1}{Z(x,y+1)}$ to evaluate the accuracy of our proposed SNE. In addition, we also compare it with two well-known surface normal estimation approaches: SRI [2] and LINE-MOD [22]. The qualitative and quantitative comparisons are shown in Fig. 3. It can be observed that our proposed SNE outperforms SRI and LINE-MOD for both indoor and outdoor scenarios.

4.3 Performance Evaluation of Our SNE-RoadSeg

In this subsection, we evaluate the performance of our proposed SNE-RoadSeg-152 (abbreviated as SNE-RoadSeg) both qualitatively and quantitatively. Ex-



(a)

(b)

Fig. 4: Examples of the experimental results on (a) the SYNTHIA road dataset and (b) our R2D road dataset: (i) RGB, (ii) Depth, (iii) SNE-Depth (Ours), (iv) RGBD and (v) SNE-RGBD (Ours); (1) DeepLabv3+ [9], (2) U-Net [26], (3) SegNet [3], (4) GSCNN [30], (5) DUpsampling [32], (6) DenseASPP [37], (7) FuseNet [19], (8) RTFNet [29], (9) Depth-aware CNN [35], (10) MFNet [18] and (11) RoadSeg (Ours). The true positive, false negative and false positive pixels are shown in green, red and blue, respectively.

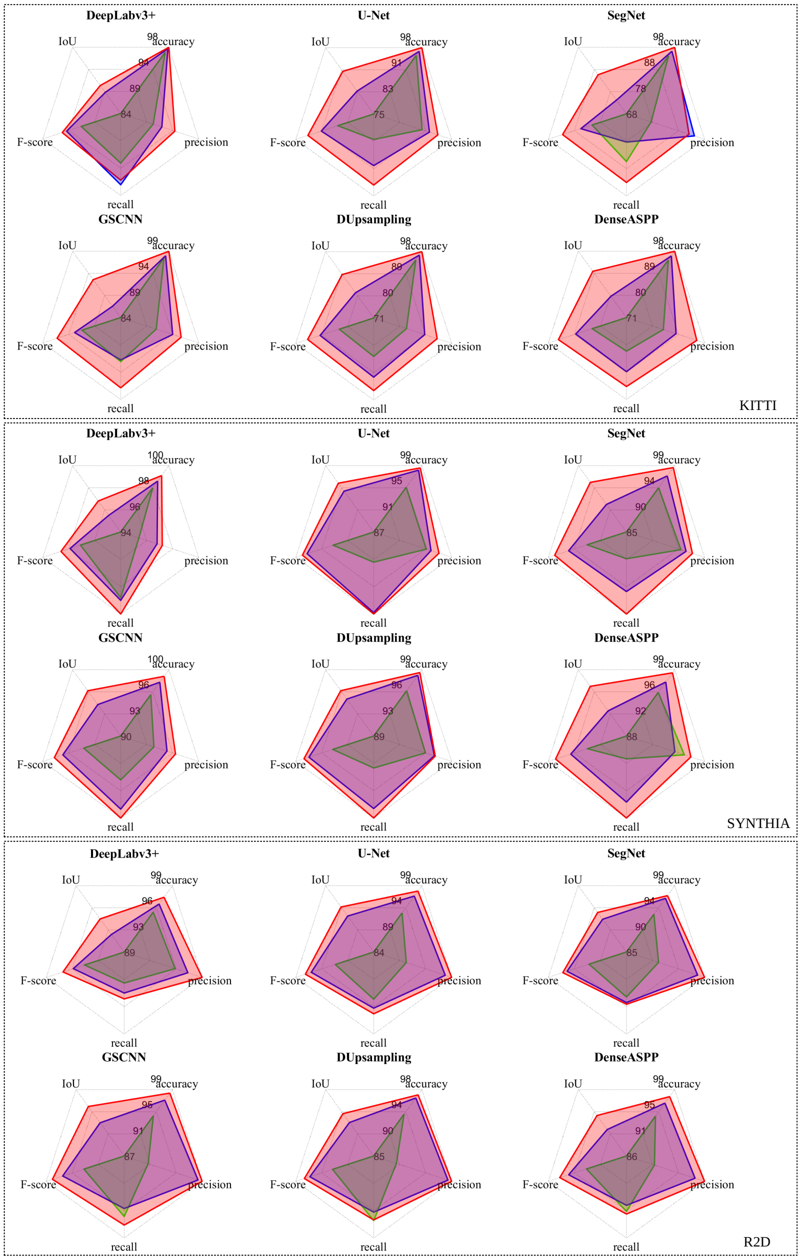


Fig. 5: Performance comparison (%) among DeepLabv3+ [9], U-Net [26], SegNet [3], GSCNN [30], DUpsampling [32] and DenseASPP [37] with and without our SNE embedded, where — RGB, — Depth, and — SNE-Depth (Ours).

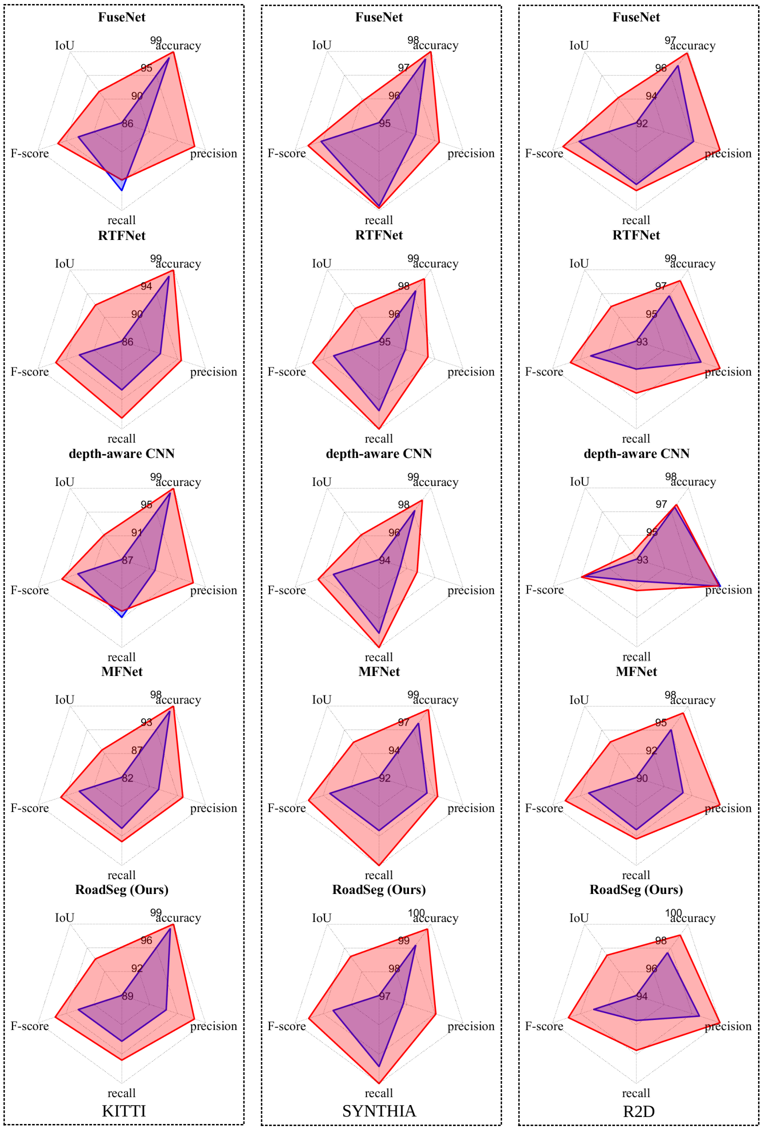


Fig. 6: Performance comparison (%) among FuseNet [19], RTFNet [29], depth-aware CNN [35], MFNet [18] and our RoadSeg with and without our SNE embedded, where — RGBD and — SNE-RGBD (Ours).

amples of the experimental results on the SYNTHIA road dataset [21] and our R2D road dataset are shown in Fig. 4. We can clearly observe that the CNNs with RGB images as inputs suffer greatly from poor illumination conditions. Moreover, the CNNs with our SNE embedded generally perform better than

Table 1: The KITTI road benchmark results, where the best results are in bold type. Please note that we only compare our method with published works.

Method	MaxF (%)	AP (%)	PRE (%)	REC (%)	Rank
RBNet [10]	93.21	89.18	92.81	93.60	21
TVFNet [17]	95.34	90.26	95.73	94.94	16
LC-CRF [16]	95.68	88.34	93.62	**97.83**	13
LidCamNet [5]	96.03	93.93	96.23	95.83	7
RBANet [28]	96.30	89.72	95.14	97.50	6
SNE-RoadSeg (Ours)	**96.75**	**94.07**	**96.90**	96.61	**2**

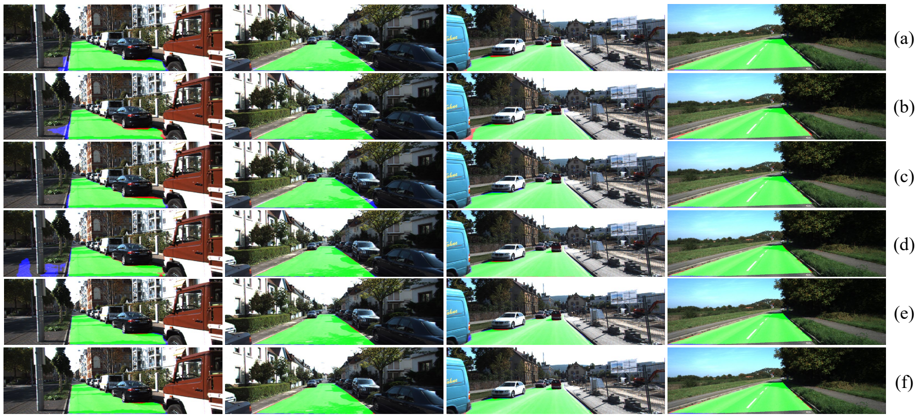


Fig. 7: Examples on the KITTI road benchmark, where rows (a)–(f) show the freespace detection results obtained by RBNet [10], TVFNet [17], LC-CRF [16], LidCamNet [5], RBANet [28] and our proposed SNE-RoadSeg, respectively. The true positive, false negative and false positive pixels are shown in green, red and blue, respectively.

they do without our SNE embedded. The corresponding quantitative comparisons are given in Fig 5 and Fig. 6. Readers can see that the IoU increases by approximately 2-12% for single-modal CNNs and by about 1-7% for data-fusion CNNs, while the F-score increases by around 1-7% for single-modal CNNs and by about 1-4% for data-fusion CNNs. We demonstrate that our proposed SNE can make the road areas become highly distinguishable, and thus, it will benefit all state-of-the-art CNNs for freespace detection.

Furthermore, from Fig 5 and Fig. 6, we can observe that RoadSeg itself outperforms all other CNNs. We demonstrate that the densely-connected skip connections utilized in our proposed RoadSeg can help achieve flexible feature fusion and smooth the gradient flow to generate accurate freespace detection results. Also, RoadSeg with our SNE embedded performs better than all other CNNs with our SNE embedded. An increase of approximately 1.4-14.7% is witnessed on the IoU, while the F-score increases by about 0.7-8.8%.

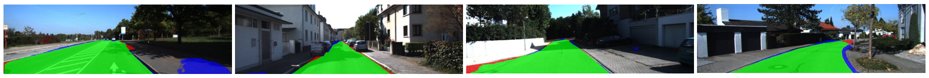


Fig. 8: Unsatisfactory results obtained using the KITTI road dataset. The true positive, false negative and false positive pixels are shown in green, red and blue, respectively.

In addition, we compare our proposed method with five state-of-the-art CNNs published on the KITTI road benchmark [15]. Examples of the experimental results are shown in Fig. 7. The quantitative comparisons are given in Table 1, which shows that our proposed SNE-RoadSeg achieves the highest MaxF (maximum F-score), AP (average precision) and PRE (precision), while LC-CRF [16] achieves the best REC (recall). Our freespace detection method is the second best on the KITTI road benchmark [15].

Fig. 8 presents several unsatisfactory results of our SNE-RoadSeg on the KITTI road dataset [15]. Since the 3D points on freespace and sidewalks possess very similar surface normals, our proposed approach can sometimes mistakenly recognize part of sidewalks as freespace, especially when the textures of the road and sidewalks are similar. We believe this can be improved by leveraging surface normal gradient features, as there usually exists a clear boundary between freespace and sidewalks (due to their differences in height).

4.4 Ablation Study

In this subsection, we conduct ablation studies on our R2D road dataset to validate the effectiveness of the architecture for our RoadSeg. The performances of different architectures are provided in Table 2.

Firstly, we replace the backbone of RoadSeg with different ResNet architectures. The quantitative results are given in Table 2. The superior performance of our choice is as expected, because ResNet-152 has also presented the best image classification performance among the five ResNet architectures [20].

Then, we remove one encoder from RoadSeg to evaluate its performance on single-modal vision data. We conduct five experiments: a) training with RGB images, denoted as **RGB**; b) training with depth images, denoted as **Depth**; c) training with depth images, denoted as **SNE-Depth**; d) training with four-channel RGB-D vision data, denoted as **RGBD-C**; and e) training with four-channel RGB-D vision data, denoted as **SNE-RGBD-C**. From Table 2, we can observe that our choice outperforms the single-modal architecture with respect to different modalities of training data, proving that the data fusion via a two-encoder architecture can benefit the freespace detection. It should be noted that although the single-modal architectures cannot provide competitive results, our proposed SNE still benefits them for better freespace detection performance.

To further validate the effectiveness of our choice, we replace the densely-connected skip connections in the decoder with two different architectures: a) no skip connections (NSCs), which totally removes the skip connections; b) sparse

Table 2: Performance comparison (%) among different architectures and setups on our R2D road dataset. The best results are shown in bold font.

Architecture	Setup	Accuracy	Precision	Recall	F-Score	IoU
RoadSeg-18	SNE-RGBD	93.6	93.5	91.3	92.4	85.9
RoadSeg-34		95.5	96.3	93.0	94.6	89.8
RoadSeg-50		96.8	97.5	95.2	96.3	92.9
RoadSeg-101		98.0	98.2	97.1	97.6	95.4
RoadSeg-152	RGB	94.0	91.9	93.8	92.8	86.6
	Depth	96.7	97.6	94.6	96.1	92.4
	SNE-Depth	97.6	98.9	95.5	97.2	94.5
	RGBD-C	95.1	92.8	95.6	94.2	89.0
	SNE-RGBD-C	97.0	97.5	95.3	96.4	93.0
RoadSeg-152-NSCs	SNE-RGBD	97.9	98.6	96.5	97.5	95.2
RoadSeg-152-SSCs		98.2	99.0	96.8	97.9	95.9
RoadSeg-152 (Ours)	SNE-RGBD	**98.6**	**99.1**	**97.6**	**98.3**	**96.7**

skip connections (SSCs), which employs the skip connections only at the same-scale feature maps of the encoder and decoder (like U-Net). Table 2 verifies the superiority of the densely-connected skip connections, which helps to achieve flexible feature fusion and to smooth the gradient flow to generate accurate freespace detection results, as analyzed in Section 4.3.

5 Conclusion

The main contributions of this paper include: a) a module named SNE, capable of inferring surface normal information from depth/disparity images with both high precision and efficiency; b) a data-fusion CNN architecture named RoadSeg, capable of fusing both RGB and surface normal information for accurate freespace detection; and c) a publicly available synthetic dataset for semantic driving scene segmentation. To demonstrate the feasibility and effectiveness of the proposed SNE module, we embedded it into ten state-of-the-art CNNs and evaluated their performances for freespace detection. The experimental results illustrated that our introduced SNE can benefit all these CNNs for freespace detection. Furthermore, our proposed data-fusion CNN architecture RoadSeg is most compatible with our proposed SNE, and it outperforms all other CNNs when detecting drivable road regions.

Acknowledgements

This work was supported by the National Natural Science Foundation of China, under grant No. U1713211, and the Research Grant Council of Hong Kong SAR Government, China, under Project No. 11210017, awarded to Prof. Ming Liu.

References

1. Alvarez, J.M., Gevers, T., LeCun, Y., Lopez, A.M.: Road scene segmentation from a single image. In: European Conference on Computer Vision. pp. 376–389. Springer (2012)
2. Badino, H., Huber, D., Park, Y., Kanade, T.: Fast and accurate computation of surface normals from range images. In: 2011 IEEE International Conference on Robotics and Automation. pp. 3084–3091. IEEE (2011)
3. Badrinarayanan, V., Kendall, A., Cipolla, R.: Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE transactions on pattern analysis and machine intelligence **39**(12), 2481–2495 (2017)
4. Cai, P., Wang, S., Sun, Y., Liu, M.: Probabilistic end-to-end vehicle navigation in complex dynamic environments with multimodal sensor fusion. IEEE Robotics and Automation Letters **5**, 4218–4224 (2020)
5. Caltagirone, L., Bellone, M., Svensson, L., Wahde, M.: Lidar–camera fusion for road detection using fully convolutional neural networks. Robotics and Autonomous Systems **111**, 125–131 (2019)
6. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Semantic image segmentation with deep convolutional nets and fully connected crfs. CoRR **abs/1412.7062** (2014)
7. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE transactions on pattern analysis and machine intelligence **40**(4), 834–848 (2017)
8. Chen, L.C., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587 (2017)
9. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European conference on computer vision (ECCV). pp. 801–818 (2018)
10. Chen, Z., Chen, Z.: Rbnet: A deep neural network for unified road and road boundary detection. In: International Conference on Neural Information Processing. pp. 677–687. Springer (2017)
11. Dosovitskiy, A., Ros, G., Codevilla, F., Lopez, A., Koltun, V.: CARLA: An open urban driving simulator. In: Levine, S., Vanhoucke, V., Goldberg, K. (eds.) Proceedings of the 1st Annual Conference on Robot Learning. Proceedings of Machine Learning Research, vol. 78, pp. 1–16. PMLR (13–15 Nov 2017)
12. Fan, R., Jiao, J., Pan, J., Huang, H., Shen, S., Liu, M.: Real-time dense stereo embedded in a UAV for road inspection. In: Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 535–543 (2019)
13. Fan, R., Ozgunalp, U., Hosking, B., Liu, M., Pitas, I.: Pothole detection based on disparity transformation and road surface modeling. IEEE Transactions on Image Processing **29**, 897–908 (2019)
14. Fan, R., Wang, H., Xue, B., Huang, H., Wang, Y., Liu, M., Pitas, I.: Three-filters-to-normal: An accurate and ultrafast surface normal estimator. arXiv preprint arXiv:2005.08165 (2020), under peer review
15. Fritsch, J., Kuehnl, T., Geiger, A.: A new performance measure and evaluation benchmark for road detection algorithms. In: International Conference on Intelligent Transportation Systems (ITSC) (2013)
16. Gu, S., Zhang, Y., Tang, J., Yang, J., Kong, H.: Road detection through crf based lidar-camera fusion. In: 2019 International Conference on Robotics and Automation (ICRA). pp. 3832–3838. IEEE (2019)

17. Gu, S., Zhang, Y., Yang, J., Alvarez, J.M., Kong, H.: Two-view fusion based convolutional neural network for urban road detection. In: 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 6144–6149. IEEE (2019)
18. Ha, Q., Watanabe, K., Karasawa, T., Ushiku, Y., Harada, T.: Mfnet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes. In: 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 5108–5115. IEEE (2017)
19. Hazirbas, C., Ma, L., Domokos, C., Cremers, D.: Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture. In: Asian conference on computer vision. pp. 213–228. Springer (2016)
20. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
21. Hernandez-Juarez, D., Schneider, L., Espinosa, A., Vazquez, D., Lopez, A.M., Franke, U., Pollefeys, M., Moure, J.C.: Slanted stixels: Representing san francisco's steepest streets. In: British Machine Vision Conference (BMVC), 2017 (2017)
22. Hinterstoisser, S., Cagniart, C., Ilic, S., Sturm, P., Navab, N., Fua, P., Lepetit, V.: Gradient response maps for real-time detection of textureless objects. IEEE transactions on pattern analysis and machine intelligence **34**(5), 876–888 (2011)
23. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4700–4708 (2017)
24. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3431–3440 (2015)
25. Lu, C., van de Molengraft, M.J.G., Dubbelman, G.: Monocular semantic occupancy grid mapping with convolutional variational encoder–decoder networks. IEEE Robotics and Automation Letters **4**(2), 445–452 (2019)
26. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
27. Sless, L., El Shlomo, B., Cohen, G., Oron, S.: Road scene understanding by occupancy grid learning from sparse radar clusters using semantic segmentation. In: Proceedings of the IEEE International Conference on Computer Vision Workshops. pp. 0–0 (2019)
28. Sun, J.Y., Kim, S.W., Lee, S.W., Kim, Y.W., Ko, S.J.: Reverse and boundary attention network for road segmentation. In: Proceedings of the IEEE International Conference on Computer Vision Workshops. pp. 0–0 (2019)
29. Sun, Y., Zuo, W., Liu, M.: Rtfnet: Rgb-thermal fusion network for semantic segmentation of urban scenes. IEEE Robotics and Automation Letters **4**(3), 2576–2583 (2019)
30. Takikawa, T., Acuna, D., Jampani, V., Fidler, S.: Gated-scnn: Gated shape cnns for semantic segmentation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 5229–5238 (2019)
31. Thoma, J., Paudel, D.P., Chhatkuli, A., Probst, T., Gool, L.V.: Mapping, localization and path planning for image-based navigation using visual features and map. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7383–7391 (2019)
32. Tian, Z., He, T., Shen, C., Yan, Y.: Decoders matter for semantic segmentation: Data-dependent decoding enables flexible feature aggregation. In: Proceedings of

the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3126–3135 (2019)

33. Vasiljevic, I., Kolkin, N., Zhang, S., Luo, R., Wang, H., Dai, F.Z., Daniele, A.F., Mostajabi, M., Basart, S., Walter, M.R., et al.: Diode: A dense indoor and outdoor depth dataset. arXiv preprint arXiv:1908.00463 (2019)
34. Wang, H., Fan, R., Sun, Y., Liu, M.: Applying surface normal information in drivable area and road anomaly detection for ground mobile robots. In: 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (2020), to be published
35. Wang, W., Neumann, U.: Depth-aware cnn for rgb-d segmentation. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 135–150 (2018)
36. Wedel, A., Badino, H., Rabe, C., Loose, H., Franke, U., Cremers, D.: B-spline modeling of road surfaces with an application to free-space estimation. IEEE transactions on Intelligent transportation systems **10**(4), 572–583 (2009)
37. Yang, M., Yu, K., Zhang, C., Li, Z., Yang, K.: Denseaspp for semantic segmentation in street scenes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3684–3692 (2018)