SNE-RoadSeg+: Rethinking Depth-Normal Translation and Deep Supervision for Freespace Detection

Hengli Wang[1]*, Rui Fan[2]*, Peide Cai[1], and Ming Liu[1], *Senior Member, IEEE*

***Abstract*— Freespace detection is a fundamental component of autonomous driving perception. Recently, deep convolutional neural networks (DCNNs) have achieved impressive performance for this task. In particular, SNE-RoadSeg, our previously proposed method based on a surface normal estimator (SNE) and a data-fusion DCNN (RoadSeg), has achieved impressive performance in freespace detection. However, SNE-RoadSeg is computationally intensive, and it is difficult to execute in real time. To address this problem, we introduce SNE-RoadSeg+, an upgraded version of SNE-RoadSeg. SNE-RoadSeg+ consists of 1) SNE+, a module for more accurate surface normal estimation, and 2) RoadSeg+, a data-fusion DCNN that can greatly minimize the trade-off between accuracy and efficiency with the use of deep supervision. Extensive experimental results have demonstrated the effectiveness of our SNE+ for surface normal estimation and the superior performance of our SNE-RoadSeg+ over all other freespace detection approaches. Specifically, our SNE-RoadSeg+ runs in real time, and meanwhile, achieves the state-of-the-art performance on the KITTI road benchmark. Our project page is at `https://www.sne-roadseg.site/sne-roadseg-plus`.**

I. Introduction

Autonomous driving appears prominently in our society in the form of the advanced driver assistance system (ADAS) in both commercial and research vehicles [1]. Visual environment perception, the front-end module and key component of the ADAS, analyzes the raw data collected by the car's sensors and outputs its understanding to the driving scenario [2]–[4]. Its outputs are then used by other modules, such as prediction and planning, to ensure the safe navigation of self-driving cars in complex environments [5], [6].

As a fundamental task in visual environment perception, freespace detection performs pixel-level binary classification on vision sensor data, *e.g.*, RGB images [7], LiDAR point clouds [8], or depth/disparity images [9]. This is generally realized with traditional segmentation algorithms and/or deep convolutional neural networks (DCNNs) [10]. With the use of modern encoder-decoder architectures, semantic segmentation DCNNs have emerged as the most powerful tool for robust freespace detection, and their performance under different environmental conditions is incredibly good. Therefore, many researchers have turned their focuses towards developing DCNN-based freespace detection approaches.

Recent data-fusion DCNNs for semantic segmentation [11]–[14] have achieved the state-of-the-art (SOTA) performance in freespace detection by extracting visual features from different modalities of vision sensor data and fusing the extracted features to provide accurate semantic prediction. For example, progressive LiDAR adaptation-aided road detection (PLARD) [8] learns both visual and LiDAR features using two DCNNs. A feature space adaptation module then follows to adapt the LiDAR features to visual features. This helps PLARD [8] achieve impressive freespace detection results. Furthermore, we recently introduced a SOTA freespace detection algorithm, named SNE-RoadSeg [12]. It consists of 1) a surface normal estimator (SNE), a lightweight module for efficient end-to-end translation from depth/disparity images into surface normal inference maps, and 2) RoadSeg, a data-fusion DCNN capable of extracting and fusing features from both RGB images and the inferred surface normal maps for accurate freespace detection. However, SNE-RoadSeg [12] is computationally intensive, and it is difficult to execute in real time. One possible solution is to reduce the network depth/level of RoadSeg, but the selection of the optimal depth requires an extensive architecture search. Furthermore, in order to improve the computational efficiency, the SNE hypothesizes that the angle between an arbitrary pair of normalized surface normals is less than $\pi/2$ [12]. This can, sometimes, degrade the performance of the SNE near ambiguities, further deteriorating freespace detection performance.

To resolve the limitations of SNE-RoadSeg, in this paper, we introduce SNE-RoadSeg+, an upgraded version of SNE-RoadSeg, as shown in Fig. 1. SNE-RoadSeg+ consists of 1) SNE+, a module for surface normal information inference without the hypothesis made in [12], and 2) RoadSeg+, a data-fusion DCNN that can greatly minimize the trade-off between accuracy and efficiency with the use of deep supervision. Extensive experimental results on the DIODE [15] and ScanNet [16] datasets have demonstrated the effectiveness of our SNE+ for surface normal inference. Moreover, we have evaluated our SNE-RoadSeg+ on the KITTI road [17] and Ready-to-Drive (R2D) road [12] datasets. The achieved results show that our SNE-RoadSeg+ outperforms all other freespace detection approaches. Specifically, SNE-RoadSeg+

This work was supported in part by the Zhongshan Municipal Science and Technology Bureau Fund under Project ZSST21EG06, in part by the Collaborative Research Fund by Research Grants Council Hong Kong under Project C4063-18G, and in part by the Department of Science and Technology of Guangdong Province Fund under Project GDST20EG54, awarded to Prof. Ming Liu. *(Corresponding author: Ming Liu.)*

[1]H. Wang, P. Cai and M. Liu are with the Department of Electronic and Computer Engineering, the Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong SAR, China (email: {hwangdf, peide.cai, eelium}@ust.hk).

[2]R. Fan is with the College of Electronic and Information Engineering, Tongji University, Shanghai 201804, P. R. China, as well as Shanghai Research Institute for Intelligent Autonomous Systems, Shanghai 201210, P. R. China (email: rui.fan@ieee.org).

*The authors contributed equally to this work.

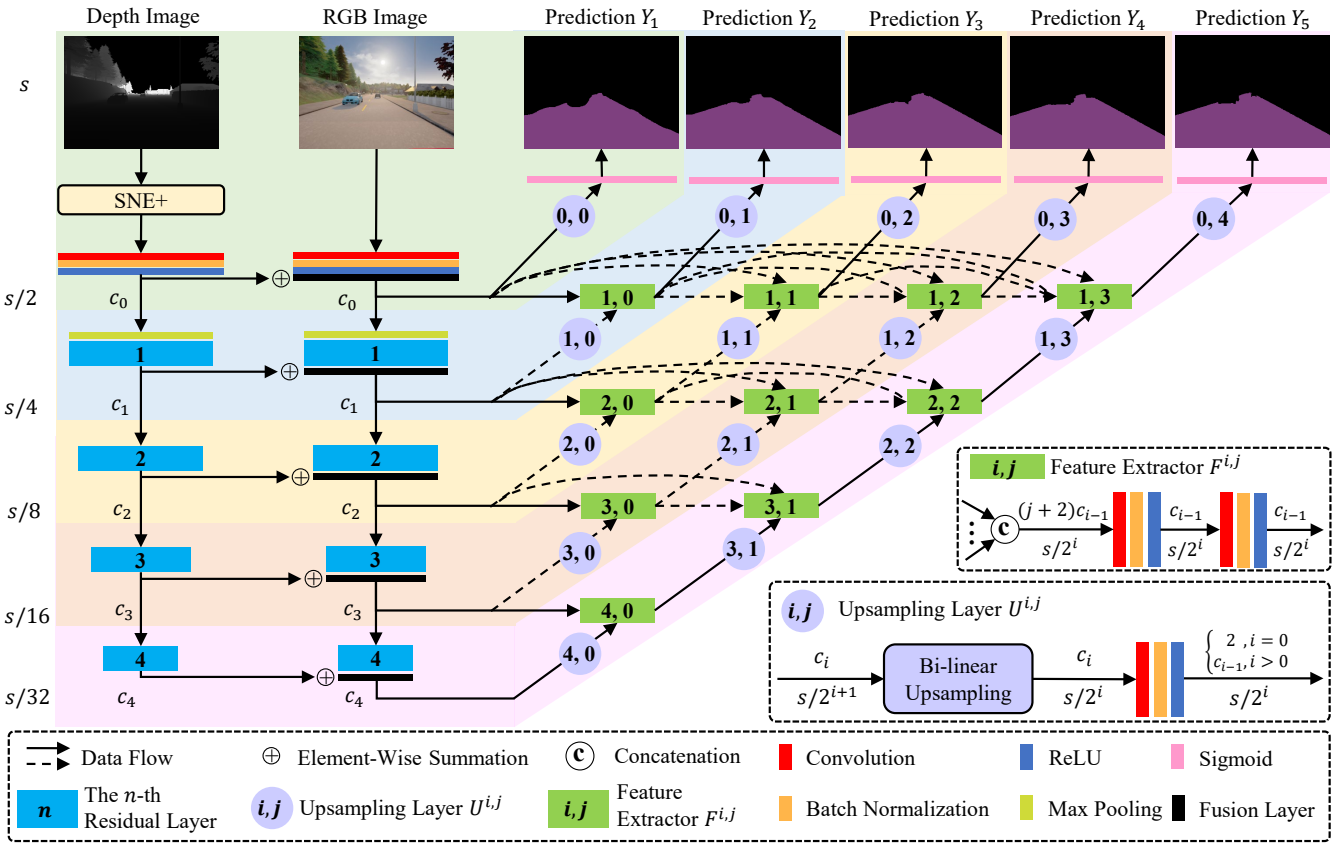

Fig. 1: An overview of our proposed freespace detection framework SNE-RoadSeg+. It consists of 1) SNE+, a lightweight module for accurate surface normal information inference, and 2) RoadSeg+, a data-fusion DCNN that can greatly minimize the trade-off between accuracy and efficiency with the use of deep supervision. The entire framework has five network levels, which are illustrated in different background colors. During the inference phase, we can either 1) assemble all five freespace predictions to achieve the most accurate performance or 2) prune the network and only use the low-level parts to achieve the required efficiency without additional training. s denotes the resolution of the input RGB and depth images, and c_n denotes the number of feature map channels at different levels.

runs in real time, and meanwhile, achieves the state-of-the-art performance on the KITTI road benchmark[1] [17].

II. Related Work

A. Surface Normal Estimation

Existing surface normal estimation methods can be basically categorized as data-driven [18], [19] or geometry-based [20], [21]. The former algorithms are generally trained using supervised learning techniques, hence requiring a large amount of well-labeled surface normal ground truth to find the best network parameters. Moreover, such algorithms are not designed specifically for surface normal estimation, because they are only considered as auxiliary functionalities for other computer vision and robotics tasks. Prior to our works [12], [22], geometry-based algorithms typically fit local (planar or quadratic) surfaces to a collection of 3D points adjacent to the observed 3D point. By minimizing residual error using optimization techniques, such as principal component analysis or singular value decomposition, the surface normal information can be obtained. Recently, we proposed three-filters-to-normal (3F2N) [22] and SNE-RoadSeg [12], which provide an efficient way to translate depth/disparity images directly into accurate surface normal maps. In this paper, we introduce SNE+, a more accurate surface normal estimation approach, which neglects the hypothesis made in SNE-RoadSeg [12]. In addition, our SNE+ can also benefit DCNNs in freespace detection, as will be demonstrated in Section V.

B. Semantic Segmentation DCNNs

Since [23] proposed the fully convolutional network (FCN), many single-modal networks have been developed for semantic segmentation. SegNet [24] first presented the popular encoder-decoder architecture, which is widely used in current networks. U-Net [25] adopts the encoder-decoder paradigm, and further incorporates skip connections into the network for performance improvement. Moreover, DeepLabv3+ [26] and DenseASPP [27] combine the advantages of the spatial pyramid pooling (SPP) module and the encoder-decoder architecture to improve the semantic prediction detail. In addition, GSCNN [28] utilizes the boundary information to refine the semantic predictions.

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

To further improve the semantic segmentation performance, some researchers have developed data-fusion networks that use two (or more) types of visual features. Specifically, FuseNet [29] utilizes RGB images and depth images based on the popular encoder-decoder architecture. Similarly, Depth-aware CNN [30] presents two novel operations, depth-aware convolution and depth-aware average pooling, to extract useful information from depth images for performance improvement. Moreover, RTFNet [31] was designed to perform semantic segmentation using RGB images and thermal images. Additionally, our SNE-RoadSeg [12] first transforms the depth/disparity images into surface normal maps, and then fuses the features learned from both RGB images and the inferred surface normal maps for accurate freespace detection. However, SNE-RoadSeg [12] is computationally intensive, and it is difficult to execute in real time. SNE-RoadSeg+ is, thus, proposed to address this problem, and it can greatly minimize the trade-off between accuracy and efficiency with the use of deep supervision.

C. Deep Supervision

Deep supervision aims at improving the network performance by providing supervision on the intermediate layers of the network [32]. This paradigm has been used in many tasks, such as semantic segmentation. For example, the architecture design of DenseASPP [27] adopts the concept of deep supervision implicitly, while [33] added additional supervision layers to improve the performance of semantic segmentation. Following these studies, we also incorporate deep supervision into SNE-RoadSeg+ for accurate and efficient freespace detection, making SNE-RoadSeg+ the first data-fusion DCNN to adopt the deep supervision paradigm. Moreover, we follow [33] and adopt a model pruning approach to achieve a great trade-off between accuracy and efficiency using deep supervision.

III. SNE+

We demonstrated in [12] and [22] that the surface normal information can be accurately and efficiently inferred from dense depth/disparity images in an end-to-end manner. These approaches first estimate the gradient $\mathbf{g}_{xy} = (n_x; n_y)$ of the surface normal's projection on the xy-plane by performing gradient filtering on a given disparity (or inverse depth) image [12]. The preliminaries of $\mathbf{g}_{xy}$ estimation are given in the supplement. Given a 3D point $\mathbf{p}$ adjacent to the observed 3D point $\mathbf{q}$, an n_z candidate can be obtained. As all candidates share one $\mathbf{g}_{xy}$, their provided surface normals are on the same tangent spherical surface. Therefore, the estimation of $\hat{n}$ (the optimum n) can be realized by finding a point on the arc of this tangent spherical surface where the n_z candidates are distributed most intensively. The key to designing an SNE thus turns into the way of formulating

$$\hat{n}_z = \Phi(\mathbf{g}_{xy}, \mathbf{q}, \mathscr{P}), \tag{1}$$

where $\mathscr{P} = (\mathbf{p}_1; \mathbf{p}_2; \ldots; \mathbf{p}_k)$ is a group of k neighboring 3D points around the observed 3D point $\mathbf{q}$. [22] formulates (1) as a median or mean filter. The former achieves better accuracy, but meanwhile, it is more computationally intensive because of the sorting operation [22]. Furthermore, SNE-RoadSeg [12] formulates (1) as an energy minimization problem with respect to inclination and azimuth. However, it hypothesizes that the angle between an arbitrary pair of surface normals is less than $\pi/2$, limiting its performance near ambiguities, where the surface normal candidates can differ significantly from each other. Since a surface normal is undirected ($\mathbf{n}$ and $-\mathbf{n}$ can be considered to be exactly the same), we formulate (1) as follows:

$$\hat{n}_z = \Big[\sin\theta\cos\varphi, \sin\theta\sin\varphi, \cos\theta\Big]^\top, \tag{2}$$

where

$$\begin{aligned}\theta &= \arg\max_{\theta} \sum_{i=1}^{k} (A_i \sin\theta + n_{z_i}\cos\theta)^2 \\ &= \frac{1}{2}\arctan\left(\frac{2\sum_{i=1}^{k} A_i n_{z_i}}{\sum_{i=1}^{k}\left({n_{z_i}}^2 - {A_i}^2\right)}\right) + \frac{\pi}{2} l, l \in \{0,1\}\end{aligned} \tag{3}$$

$\in [0, \pi]$ denotes inclination, $\varphi \in [0, 2\pi)$ denotes azimuth [12], and $A_i = n_{x_i}\cos\varphi + n_{y_i}\sin\varphi$. Compared with the SNE used in SNE-RoadSeg [12], (2) can produce a more accurate $\hat{n}_z$, and therefore, we refer to it as SNE+.

IV. RoadSeg+

Based on our previous work SNE-RoadSeg [12], our SNE-RoadSeg+ first employs the above-mentioned SNE+ to translate depth/disparity images into surface normal maps, and then fuses the features learned from both RGB images and the inferred surface normal maps for accurate freespace detection, as shown in Fig. 1. The data-fusion DCNN follows the popular encoder-decoder paradigm. Specifically, the encoder first extracts and fuses the different modalities of features in a multi-scale fashion. The decoder then utilizes feature extractors $F^{i,j}$ and upsampling layers $U^{i,j}$ to realize flexible feature fusion and accurate freespace detection. Readers are recommended to refer to [12] for more details on the network architecture. The rest of this section mainly introduces the major difference between SNE-RoadSeg [12] and SNE-RoadSeg+, that is, incorporating deep supervision into the network to improve the accuracy and efficiency for freespace detection.

Different from SNE-RoadSeg [12], which only uses one upsampling layer $U^{0,4}$ to perform the final freespace prediction, we append four upsampling layers, namely, $U^{0,0}$, $U^{0,1}$, $U^{0,2}$ and $U^{0,3}$, to output freespace predictions at different network levels/depths. During the training phase, the freespace prediction Y_i at level i is supervised by the cross entropy loss as follows:

$$\mathcal{L}_i\left(\widehat{Y}, Y_i\right) = -\sum_{\mathbf{p}} \widehat{Y}\left(\mathbf{p}\right) \cdot \log\left(Y_i\left(\mathbf{p}\right)\right), \tag{4}$$

where $\mathbf{p}$ denotes the valid pixels and $\widehat{Y}$ denotes the freespace ground truth. The adopted overall loss $\mathcal{L}$ is then defined as a weighted summation of $\mathcal{L}_i$, *i.e.*, $\mathcal{L} = \sum_{i=1}^{5} \alpha_i \mathcal{L}_i$. During the inference phase, we can assemble the freespace estimations

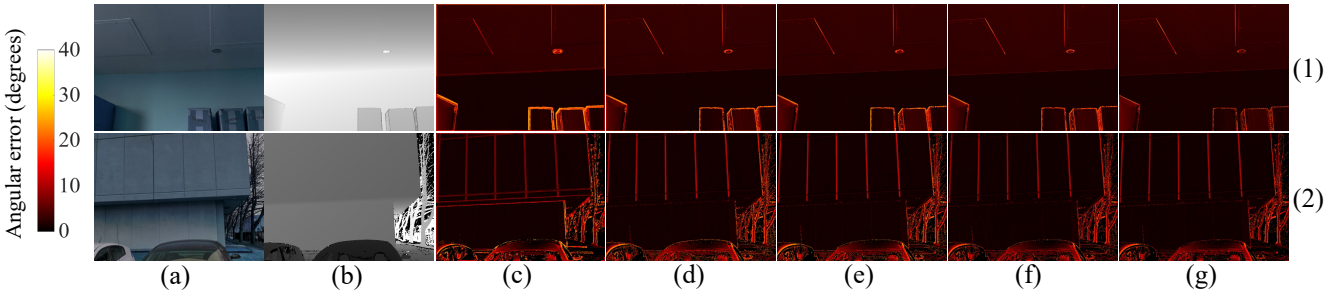

Fig. 2: Examples of the surface normal estimation results on the DIODE dataset [15]: (a) RGB images; (b) depth images; (c)–(g) angular error maps of LINE-MOD [20], PlanePCA [21], 3F2N [22], SNE [12] and our SNE+, respectively; (1) the indoor scenario; and (2) the outdoor scenario.

at all five network levels by computing their average to generate the final freespace prediction.

Additionally, we follow [33] and adopt a model pruning approach to optimize the trade-off between accuracy and efficiency based on the deep supervision paradigm. Specifically, since the freespace predictions at low network levels do not rely on the high-level network architecture, we can prune arbitrary high-level layers of the architecture to achieve flexible acceleration with acceptable performance degradation. For example, we can adopt the network with only two levels, *i.e.*, the green and blue parts shown in Fig. 1, and the final freespace estimation can be obtained by assembling Y_1 and Y_2. Please note that this model pruning process can be directly conducted during the inference phase to achieve the required efficiency without additional training.

In summary, the advantages of the proposed deep supervision paradigm are twofold. 1) Providing supervision on the intermediate layers of the network can smooth the gradient flow for effective and efficient training, further leading to accurate freespace prediction results. 2) The deep supervision paradigm ensures that the intermediate layers of the network can provide accurate freespace predictions. Therefore, we do not require additional training after pruning the network. Instead, we can directly use the pruned network during the inference phase to boost network efficiency.

V. Experimental Results and Discussion

A. Datasets and Experimental Setups

Two datasets are adopted to evaluate the performance of our SNE+ for surface normal estimation:

- The DIODE dataset [15]: This dataset provides approximately 25K depth images with the corresponding surface normal ground truth in real-world indoor and outdoor scenarios.
- The ScanNet dataset [16]: This dataset contains about 2.5M depth images with the corresponding surface normal ground truth in 1513 real-world indoor scenarios.

Moreover, two other datasets are used to evaluate the performance of our SNE-RoadSeg+ for freespace detection:

- The KITTI road dataset [17]: This dataset provides RGB-D data in real-world driving scenarios. Specifically, it contains 289 paris of training data with ground truth for freespace detection and 290 pairs of testing data without ground truth.
- The R2D road dataset [12]: This dataset is a synthetic dataset collected under different illumination and weather conditions. It contains 11430 pairs of RGB-D images with the corresponding ground truth for freespace detection.

In our experiments, we first compare our SNE+ with state-of-the-art surface normal estimation approaches, as presented in Section V-B. Then, we compare our SNE-RoadSeg+ with state-of-the-art DCNNs for freespace detection. Specifically, we adopt the stochastic gradient descent with momentum (SGDM) optimizer during the training phase. We also adopt the early stopping mechanism to avoid over-fitting. The corresponding experimental results are presented in Section V-C. Finally, we submit the results achieved by our SNE-RoadSeg+ to the KITTI road benchmark [17], as presented in Section V-D.

In addition, we adopt the average angular error e_A to quantify the performance of surface normal estimation approaches:

$$e_A = \frac{1}{m}\sum_{k=1}^{m}\cos^{-1}\left(\frac{\langle \mathbf{n}_k, \hat{\mathbf{n}}_k\rangle}{\|\mathbf{n}_k\|_2 \|\hat{\mathbf{n}}_k\|_2}\right), \tag{5}$$

where m denotes the number of valid (observed) pixels; and $\mathbf{n}_k$ and $\hat{\mathbf{n}}_k$ denote the ground-truth and estimated surface normals, respectively. An accurate surface normal estimation approach achieves a low e_A value. Furthermore, two commonly used metrics are adopted for the performance evaluation of freespace detection, namely, the F-score (Fsc) and intersection over union (IoU):

$$\text{Fsc} = \frac{2n_{\text{tp}}^2}{2n_{\text{tp}}^2 + n_{\text{tp}}\left(n_{\text{fp}} + n_{\text{fn}}\right)} \times 100\%, \tag{6}$$

$$\text{IoU} = \frac{n_{\text{tp}}}{n_{\text{tp}} + n_{\text{fp}} + n_{\text{fn}}} \times 100\%, \tag{7}$$

where n_{tp}, n_{tn}, n_{fp} and n_{fn} denote the true positive, true negative, false positive and false negative pixel numbers, respectively. An accurate freespace detection approach achieves high Fsc and IoU values.

B. Evaluations for Surface Normal Estimation

We compare our SNE+ with four SOTA surface normal estimation approaches: LINE-MOD [20], PlanePCA [21],

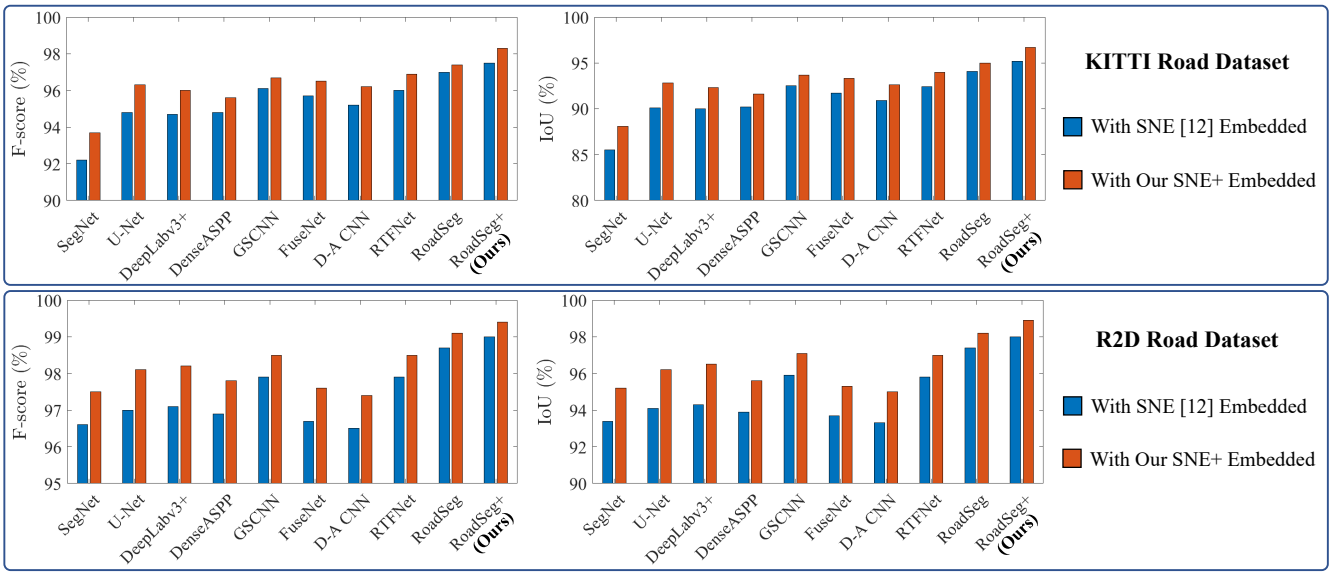

Fig. 3: DCNN comparison for freespace detection performance on the KITTI road [17] and R2D road [12] datasets. "D-A CNN" is the abbreviation of "Depth-aware CNN". SegNet [24], U-Net [25], DeepLabv3+ [26], DenseASPP [27] and GSCNN [28] are single-modal DCNNs, while FuseNet [29], Depth-aware CNN [30], RTFNet [31], RoadSeg [12] and our RoadSeg+ are data-fusion DCNNs.

TABLE I: e_A (degrees) of surface normal estimation approaches on the DIODE [15] and ScanNet [16] datasets. The best results are shown in bold type.

Approach	DIODE		ScanNet
	Indoor	Outdoor	
LINE-MOD [20]	12.839	17.272	14.479
PlanePCA [21]	10.888	16.579	13.164
3F2N [22]	10.589	16.254	12.628
SNE [12]	10.316	15.431	12.669
SNE+ (**Ours**)	**10.205**	**15.136**	**12.373**

3F2N [22] and SNE [12]. The quantitative results are presented in Table I, where it can be seen that our SNE+ outperforms all other approaches on both the DIODE [15] and ScanNet [16] datasets. Examples of the qualitative results are shown in Fig. 2, where it can be observed that our SNE+ performs better near object boundaries. This is due to the ability of our proposed parameterization method to greatly minimize the effects caused by ambiguities.

C. Evaluations for Freespace Detection

Since the whole framework shown in Fig. 1 has five network levels, we can use the proposed model pruning approach to generate networks with different network depths. Considering the trade-off between accuracy and efficiency, we use the network with three levels in the rest of our experiments, and hereafter refer to it as RoadSeg+. Since we demonstrated in [12] that using surface normal information can effectively improve the freespace detection performance, we now focus on verifying the superiorities of 1) our SNE+ over SNE [12], and 2) our RoadSeg+ over SOTA DCNNs for freespace detection. Specifically, each single-modal DCNN takes depth images as input, and each data-fusion DCNN takes RGB and depth images as input. In addition, each DCNN is evaluated with SNE [12] embedded and with our SNE+ embedded, respectively. The corresponding quantitative results are presented in Fig. 3. It is observed that the DCNNs with our SNE+ embedded outperform themselves with SNE [12] embedded. Moreover, our RoadSeg+ with our SNE+ embedded performs better than all other DCNNs, with an IoU increment of around 1–11%. Qualitative results are provided in the supplement. All these results strongly prove the effectiveness of the proposed framework, which we refer to as SNE-RoadSeg+. We next submit its results to the KITTI road benchmark [17], as presented in Section V-D.

TABLE II: Results on the KITTI road benchmark [17], where the best results are shown in bold type.

Approach	MaxF (%)	AP (%)	Runtime (s)
RBNet [34]	94.97	91.49	0.18
LC-CRF [35]	95.68	88.34	0.18
LidCamNet [36]	96.03	93.93	0.15
SNE-RoadSeg [12]	96.75	**94.07**	0.10
PLARD [8]	97.03	94.03	0.16
SNE-RoadSeg+ (**Ours**)	**97.50**	93.98	**0.08**

D. Performance on the KITTI Road Benchmark

Table II presents the KITTI road benchmark [17] results, where our SNE-RoadSeg+ achieves the state-of-the-art performance with a real-time inference speed. Excitingly, our SNE-RoadSeg+ outperforms all other freespace detection approaches in terms of both accuracy and efficiency. Fig. 4 illustrates an example of the testing images on the benchmark, where we can observe that our SNE-RoadSeg+ can present more accurate freespace detection estimations. By accelerating the proposed SNE-RoadSeg+ with TensorRT, it can run in real time on resource-limited embedded computing

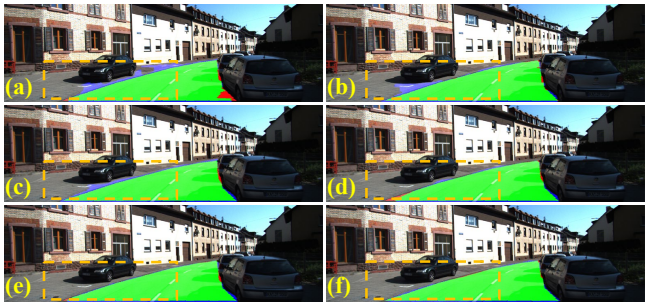

Fig. 4: An example of the testing images on the KITTI road benchmark [17]: (a) RBNet [34]; (b) LC-CRF [35]; (c) LidCamNet [36]; (d) SNE-RoadSeg [12]; (e) PLARD [8]; and (f) our SNE-RoadSeg+. Green, blue and red pixels correspond to the true positives, false positives and false negatives, respectively. Significantly improved regions are marked with orange dashed boxes.

platforms. Therefore, SNE-RoadSeg+ is more capable than SNE-RoadSeg for practical autonomous driving applications.

VI. Conclusion

In this paper, we proposed SNE-RoadSeg+, an effective and efficient approach for freespace detection. Our SNE-RoadSeg+ consists of 1) SNE+, a lightweight module for accurate surface normal estimation, and 2) RoadSeg+, a data-fusion DCNN that can achieve a great trade-off between accuracy and efficiency with the use of deep supervision. Extensive experimental results demonstrated 1) the effectiveness of our SNE+ for surface normal estimation, and 2) the superior performance of our SNE-RoadSeg+ over all other SOTA freespace detection approaches. Specifically, our SNE-RoadSeg+ achieves the state-of-the-art performance on the KITTI road benchmark, with a real-time inference speed.

References

[1] J. Van Brummelen, M. O'Brien, D. Gruyer, and H. Najjaran, "Autonomous vehicle perception: The technology of today and tomorrow," *Transp. Res. part C: Emerg. Technol.*, vol. 89, pp. 384–406, 2018.

[2] C. Yan *et al.*, "Supervised hash coding with deep neural network for environment perception of intelligent vehicles," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 1, pp. 284–295, 2017.

[3] H. Wang, R. Fan, P. Cai, and M. Liu, "PVStereo: Pyramid voting module for end-to-end self-supervised stereo matching," *IEEE Robot. Autom. Lett.*, vol. 6, no. 3, pp. 4353–4360, 2021.

[4] H. Wang, R. Fan, and M. Liu, "CoT-AMFlow: Adaptive modulation network with co-teaching strategy for unsupervised optical flow estimation," in *Conf. Robot Learn. (CoRL)*, 2020.

[5] H. Wang *et al.*, "Learning interpretable end-to-end vision-based motion planning for autonomous driving with optical flow distillation," in *IEEE Inter. Conf. Robot. Autom.*, 2021.

[6] H. Wang *et al.*, "End-to-end interactive prediction and planning with optical flow distillation for autonomous driving," in *IEEE/CVF Conf. Comput. Vision Pattern Recognit. Workshops*, 2021.

[7] X. Han *et al.*, "Semisupervised and weakly supervised road detection based on generative adversarial networks," *IEEE Signal Process. Lett.*, vol. 25, no. 4, pp. 551–555, 2018.

[8] Z. Chen *et al.*, "Progressive lidar adaptation for road detection," *IEEE/CAA J. Automatica Sinica*, vol. 6, no. 3, pp. 693–702, 2019.

[9] 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, 2019.

[10] R. Fan *et al.*, "Learning collision-free space detection from stereo images: Homography matrix brings better data augmentation," *IEEE/ASME Tran. Mechatronics*, 2021.

[11] H. Wang *et al.*, "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.

[12] R. Fan *et al.*, "SNE-RoadSeg: Incorporating surface normal information into semantic segmentation for accurate freespace detection," in *Eur. Conf. Comput Vis. (ECCV)*. Springer, 2020, pp. 340–356.

[13] H. Wang *et al.*, "Applying surface normal information in drivable area and road anomaly detection for ground mobile robots," in *IEEE/RSJ Inter. Conf. Intell. Robots Syst. (IROS)*. IEEE, 2020, pp. 2706–2711.

[14] 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.

[15] I. Vasiljevic *et al.*, "DIODE: A Dense Indoor and Outdoor DEpth Dataset," *CoRR*, vol. abs/1908.00463, 2019. [Online]. Available: http://arxiv.org/abs/1908.00463

[16] A. Dai *et al.*, "Scannet: Richly-annotated 3d reconstructions of indoor scenes," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2017.

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

[18] A. Bansal, B. Russell, and A. Gupta, "Marr revisited: 2d-3d alignment via surface normal prediction," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 5965–5974.

[19] J. Huang *et al.*, "Framenet: Learning local canonical frames of 3d surfaces from a single rgb image," in *Proc. IEEE/CVF Inter. Conf. Comput. Vision*, 2019, pp. 8638–8647.

[20] S. Hinterstoisser *et al.*, "Gradient response maps for real-time detection of textureless objects," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 5, pp. 876–888, 2011.

[21] K. Jordan and P. Mordohai, "A quantitative evaluation of surface normal estimation in point clouds," in *IEEE/RSJ Inter. Conf. Intell. Robots Syst.* IEEE, 2014, pp. 4220–4226.

[22] R. Fan *et al.*, "Three-filters-to-normal: An accurate and ultrafast surface normal estimator," *IEEE Robot. Autom. Lett.*, vol. 6, no. 3, pp. 5405–5412, 2021.

[23] 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.

[24] V. Badrinarayanan *et al.*, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, 2017.

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

[26] L.-C. Chen *et al.*, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Eur. Conf. Comput Vis. (ECCV)*, 2018, pp. 801–818.

[27] 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.

[28] 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.

[29] C. Hazirbas *et al.*, "Fusenet: Incorporating depth into semantic segmentation via fusion-based CNN architecture," in *Asian Conf. Comput. Vision*. Springer, 2016, pp. 213–228.

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

[31] 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, 2019.

[32] C.-Y. Lee, S. Xie, P. Gallagher, Z. Zhang, and Z. Tu, "Deeply-supervised nets," in *Artif. Intell. Statist.* PMLR, 2015, pp. 562–570.

[33] Z. Zhou *et al.*, "Unet++: Redesigning skip connections to exploit multiscale features in image segmentation," *IEEE Trans. Med. Imag.*, vol. 39, no. 6, pp. 1856–1867, 2019.

[34] Z. Chen and Z. Chen, "RBNet: A deep neural network for unified road and road boundary detection," in *Inter. Conf. Neural Inf. Process.* Springer, 2017, pp. 677–687.

[35] S. Gu *et al.*, "Road detection through crf based lidar-camera fusion," in *IEEE Inter. Conf. Robot. Automat.* IEEE, 2019, pp. 3832–3838.

[36] 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, 2019.

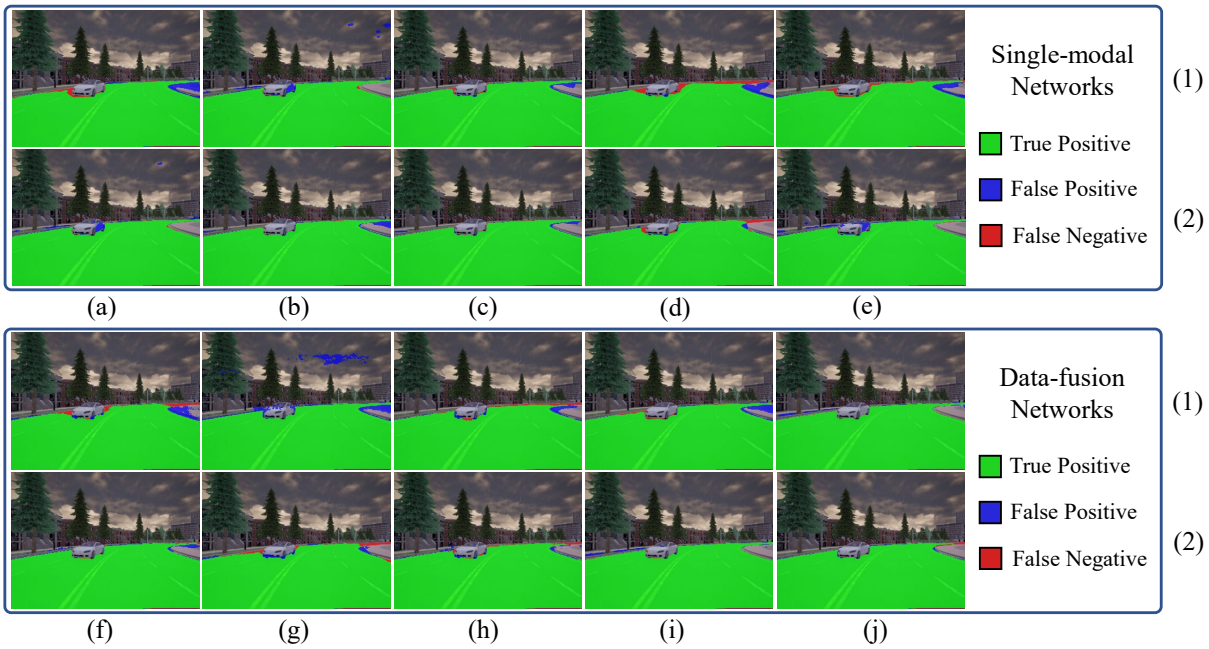

Fig. 5: Examples of the freespace detection results on the R2D road dataset [12]: (a) SegNet [24]; (b) U-Net [25]; (c) DeepLabv3+ [26]; (d) DenseASPP [27]; (e) GSCNN [28]; (f) FuseNet [29]; (g) Depth-aware CNN [30]; (h) RTFNet [31]; (i) RoadSeg [12]; (j) our RoadSeg+; (1) DCNNs with SNE [12] embedded; and (2) DCNNs with our SNE+ embedded.

SUPPLEMENT

We provide some examples of the freespace detection results in Fig. 5, where it is evident that our RoadSeg+ with our SNE+ embedded can produce more robust and accurate freespace detection results.

In basic pinhole camera models, an observed 3D point $\mathbf{q} = (x; y; z)$ can be transformed to a 2D image pixel $\mathbf{m} = (u; v)$ using $\mathbf{m} = \mathbf{K}\mathbf{q}/z$. The local planar surface $\mathcal{S}$ of $\mathbf{q}$ satisfies: $\mathbf{n}\mathbf{q} + d = 0$, where $\mathbf{n} = (n_x; n_y; n_z)$ is the surface normal of $\mathbf{q}$ and d is the distance between $\mathbf{q}$ and $\mathcal{S}$. Combining the aforementioned two equations results in:

$$1/z = -\Big((u-u_o)\cdot n_x/f_x + (v-v_o)\cdot n_y/f_y + n_z\Big)/d, \quad (8)$$

where $\mathbf{m}_o = (u_o; v_o)$ is the image principal point; and f_x and f_y are the camera focal lengths in pixels. Differentiating (8) with respect to u and v obtains:

$$n_x = -df_x\frac{\partial 1/z}{\partial u}, \quad n_y = -df_y\frac{\partial 1/z}{\partial v}. \quad (9)$$

Given an arbitrary 3D point $\mathbf{p}_i \in \mathscr{P}$ adjacent to $\mathbf{q}$, we can obtain an n_{z_i} as follows [12]:

$$n_{zi} = \frac{d}{\Delta z_i}\Big(f_x\Delta x_i\frac{\partial 1/z}{\partial u} + f_y\Delta y_i\frac{\partial 1/z}{\partial v}\Big), \quad (10)$$

where $\mathbf{r}_i = \mathbf{p}_i - \mathbf{q} = (\Delta x_i; \Delta y_i; \Delta z_i)$. As (9) and (10) have a common factor of $-d$, the expression of $\mathbf{n}_i$ (the surface normal produced by $\mathbf{q}_i$ and $\mathbf{p}$) is simplified as follows:

$$\mathbf{n}_{i_j} = \left(f_x\frac{\partial 1/z}{\partial u}; f_y\frac{\partial 1/z}{\partial v}; -\frac{f_x\Delta x_i\frac{\partial 1/z}{\partial u} + f_y\Delta y_i\frac{\partial 1/z}{\partial v}}{\Delta z_i}\right). \quad (11)$$