Fan et al. *Autonomous Intelligent Systems* (2025) 5:8
https://doi.org/10.1007/s43684-025-00096-y

Autonomous Intelligent Systems

REVIEW

Open Access

A glance over the past decade: road scene parsing towards safe and comfortable autonomous driving

Rui Fan[1*], Jiahang Li[2], Jiaqi Li[2], Jiale Wang[2], Ziwei Long[2], Ning Jia[2], Yanan Liu[3], Wenshuo Wang[4], Mohammud J. Bocus[5], Sergey Vityazev[6], Xieyuanli Chen[7], Junhao Xiao[7], Stepan Andreev[8], Huimin Lu[7] and Alexander Dvorkovich[9]

Abstract

Road scene parsing is a crucial capability for self-driving vehicles and intelligent road inspection systems. Recent research has increasingly focused on enhancing driving safety and comfort by improving the detection of both drivable areas and road defects. This article reviews state-of-the-art networks developed over the past decade for both general-purpose semantic segmentation and specialized road scene parsing tasks. It also includes extensive experimental comparisons of these networks across five public datasets. Additionally, we explore the key challenges and emerging trends in the field, aiming to guide researchers toward developing next-generation models for more effective and reliable road scene parsing.

Keywords: Road scene parsing, Self-driving vehicle, Intelligent road inspection, Drivable area, Road defect

1 Introduction

Advancements in machine intelligence and autonomous systems have dramatically fueled the integration of environmental perception technologies into daily life and various industries [1–5]. This widespread adoption is prominently seen in applications such as autonomous cars [6], smart wheelchairs [7], and unmanned ground vehicles [8]. Recently, researchers have shifted their focus toward enhancing both driving safety and comfort [9, 10]. Road scene parsing, which performs pixel-level detection of drivable areas (also known as freespace or collision-free spaces) and road defects, such as potholes and cracks, is critical for achieving these objectives [4, 11, 12].

Representative prior studies conducted in this field over the past decade are illustrated in Fig. 1. Apart from the traditional 2D image processing algorithms, such as image filtering and segmentation, which were developed decades ago, geometry-based techniques have long been the most effective approaches in this field. These techniques typically utilize explicit geometric models, such as planar and quadratic surfaces, to represent regions of interest (RoIs), which can be accurately extracted by minimizing specific energy functions. For instance, in the study [13], road surfaces are modeled as quadratic surfaces, which are interpolated from 3D point clouds through least squares fitting. A novel disparity transformation algorithm was then introduced in [14], which processes dense disparity maps to simulate a quasi-bird's eye view of the road. This transformation ensures that disparities in undamaged road areas appear uniform, making both positive and negative obstacles distinctly noticeable. Other studies, such as [15] and [16], generally employ a B-spline model to fit road disparity maps, which are subsequently projected onto a 2D v-disparity histogram for drivable area and road defect

*Correspondence: rui.fan@ieee.org
[1] College of Electronics & Information Engineering, Shanghai Institute of Intelligent Science and Technology, Shanghai Research Institute for Intelligent Autonomous Systems, the State Key Laboratory of Intelligent Autonomous Systems, and the Frontiers Science Center for Intelligent Autonomous Systems, Tongji University, 4800 Caoan Rd, Shanghai, 201804, China
Full list of author information is available at the end of the article

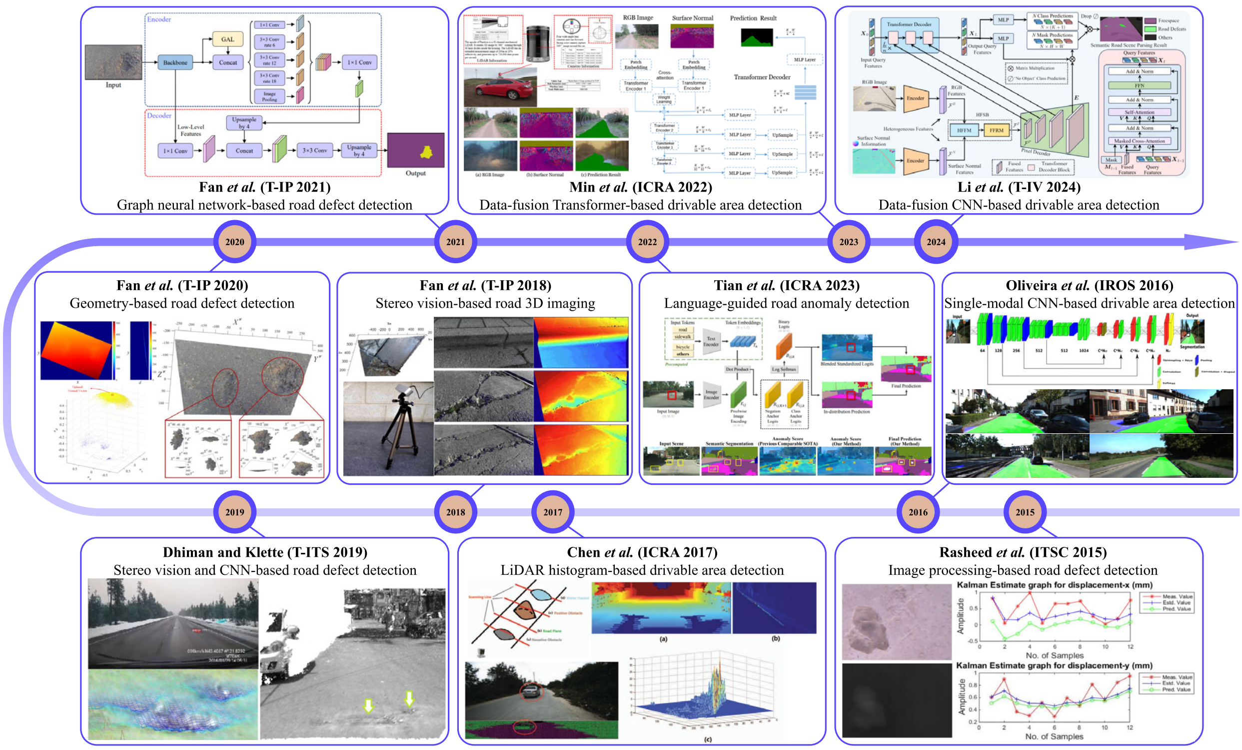

Figure 1 Representative prior studies [11, 13, 17–24] on road scene parsing conducted over the past decade

detection. However, the frequently uneven nature of actual roads can sometimes compromise the effectiveness of these methods [14].

With the advent of deep learning, convolutional neural networks (CNNs) have revolutionized road scene parsing, which is now typically formulated as a pixel-level binary or ternary classification task [11]. These methods significantly outperform traditional image processing-based and geometry-based approaches, demonstrating marked improvements in overall performance. For instance, in [25] an encoder-decoder CNN architecture is used to segment RGB images projected into a bird's eye view for drivable area detection. However, this approach often underperforms in diverse and challenging illumination and weather conditions. To address these limitations, subsequent research has explored the use of data-fusion networks with duplex encoder architectures, significantly enhancing the accuracy of road scene parsing [23, 26]. The authors of [27] extracted heterogeneous features from RGB-depth data and performed feature fusion using a straightforward element-wise summation operation. The fusion of diverse feature types brings a deeper understanding of various scenarios, resulting in superior performance over prior single-modal networks. Similarly, our SNE-RoadSeg series [6, 26, 28] achieve RGB-normal feature fusion through element-wise summation or holistic attention. By integrating a duplex backbone network with a densely-connected decoder, the SNE-RoadSeg series achieve state-of-the-art (SoTA) performance on multiple datasets, including the KITTI Road [29], vKITTI2 [30], and Cityscapes [31]. However, these approaches are constrained by either the simplistic and indiscriminate fusion of heterogeneous features or the presence of extensive trainable parameters. Both limitations can lead to conflicting feature representations and inaccurate scene parsing results.

Transformers have demonstrated their superiority over CNNs, particularly when large-scale, well-annotated datasets are available for model training [32–34]. This advantage stems from the self-attention mechanisms in Transformers, which enable more effective global context modeling compared to conventional CNNs [35]. As a result, leveraging attention mechanisms to enhance the fusion of heterogeneous features extracted by duplex encoders has become an increasingly popular research focus. OFF-Net [23] represents the first attempt to utilize the Transformer for data-fusion road scene parsing. Trained on extensive off-road datasets, it achieves marginal improvements over previous CNN-based algorithms. Nonetheless, OFF-Net employs a lightweight CNN-based decoder rather than a Transformer-based one. We believe that adopting a Transformer-based decoder could significantly enhance the upper limit of road scene parsing performance. We also observe its unsatisfactory performance in urban road scenes, particularly under data-constrained conditions [11].

Therefore, in this article, we provide a brief review of SoTA CNNs and Transformers designed for general-purpose semantic segmentation and specific road scene parsing tasks. We compare their performance across multiple public datasets to establish a benchmark for future developments. This review aims to guide readers and researchers in the field toward the creation of next-generation models for road scene parsing.

The remainder of this article is organized as follows: Sect. 2 details the most commonly used datasets for road scene parsing. Section 3 reviews SoTA networks for general semantic segmentation. Section 4 reviews networks developed specifically for road scene parsing. Section 5 presents extensive experimental results and comprehensive analyses of model performance. Section 6 discusses the existing challenges and future trends in this field of research. Finally, Sect. 7 concludes the article.

2 Existing public datasets

- **Cityscapes** [31]: This dataset is widely used for urban scene parsing. It contains 2975 stereo images for model training and 500 images for model validation, each with well-annotated semantic ground truth. All experimental results presented in this study are derived from the validation set. More details on this dataset are available at https://www.cityscapes-dataset.com. In our experiments, depth images are generated using RAFT-Stereo [36], trained on the KITTI Stereo dataset [37]. Surface normal information is subsequently computed using our proposed depth-to-normal translation algorithm [38].
- **KITTI Semantics** [39]: This dataset consists of 200 real-world RGB images, each with semantic ground-truth annotations for 19 classes, aligned with the Cityscapes dataset [31]. Detailed information is available at https://www.cvlibs.net/datasets/kitti/eval_semseg.php?benchmark=semantics2015. To evaluate the performance of road scene parsing models, semantic labels are grouped into two categories: freespace and others. Sparse disparity ground truth is obtained using a Velodyne HDL-64E LiDAR, while dense depth maps are generated using a well-trained CreStereo model [40]. In this study, the dataset is randomly split into training and validation sets in a 3:1 ratio.
- **KITTI Road** [29]: This dataset contains 289 pairs of stereo images and their corresponding LiDAR point clouds, utilized for both model training and validation. It also includes a comparable amount of testing data without semantic annotations. The model's quantitative performance on the test set has to be evaluated by submitting qualitative results to the online KITTI Road benchmark, available at https://www.cvlibs.net/datasets/kitti/eval_road.php. In this study, we adopt a data pre-processing strategy akin to that described in [26].
- **ORFD** [23]: This dataset is designed specifically for off-road drivable area detection. It contains 12,198 RGB images with corresponding LiDAR point clouds, collected across diverse scenes under different weather and illumination conditions. Additional details on this dataset can be found at https://github.com/chaytonmin/Off-Road-Freespace-Detection. In this study, we follow the data splitting and pre-processing strategies described in the original publication [23], with the exception of surface normal estimation.
- **SYN-UDTIRI** [11]: Due to the limited availability of well-annotated, large-scale datasets designed specifically for road scene parsing (including drivable area and road defect detection), the study [11] introduces a synthetic dataset called SYN-UDTIRI, developed using the CARLA simulator [41]. This dataset incorporates digital twins of real-world road potholes, created with a 3D road geometry reconstruction algorithm [14, 20]. To better replicate the roughness of actual roads, random Perlin noise is added to the road data. Six driving scenarios are simulated under different weather and illumination conditions, including rainy day, dusk, and night, as well as sunny day, dusk, and night. A simulated stereo rig with a 0.5 m baseline was mounted on a moving vehicle, generating over 10,000 pairs of stereo road images (resolution: 720×1280 pixels), along with depth images, surface normal data, and semantic annotations for three categories: drivable area, road defect, and other objects. Additional details about the SYN-UDTIRI dataset are available at https://github.com/LiJiahang617/Road-Former.

3 General semantic segmentation models for road scene parsing

This section provides a brief review of SoTA single-modal and feature-fusion general-purpose semantic segmentation networks. Their performance is evaluated both quantitatively and qualitatively in Sect. 5.

3.1 Single-modal models

Fully convolutional network (FCN) [42] marks a significant milestone in utilizing CNNs for end-to-end scene parsing. Nonetheless, its segmentation often overlooks pixel relations, leading to outputs that lack spatial consistency. Additionally, FCN incurs high memory usage and computational complexity. Fast FCN [43] addresses these issues by using upsampling convolutions to extract high-resolution feature maps, improving spatial consistency while reducing computational complexity by over threefold. Fast-SCNN [44] introduces a "learning to downsample" module for efficient computation on embedded devices with limited memory, thereby enhancing the real-time performance of semantic segmentation models on high-resolution images. Inspired by proportional-integral-derivative (PID) controllers, PIDNet [45] integrates CNNs with PID controllers, forming a novel architecture that contains three branches designed to parse detailed contextual and boundary information. PIDNet achieves an optimal balance between inference speed and accuracy.

Recent advancements in semantic segmentation have been propelled by methods that expand the receptive fields using pyramid-based multi-resolution feature extraction techniques [46–51]. For instance, DeepLabv3 [48] utilizes parallel atrous spatial pyramid pooling (ASPP) modules to capture contextual information at multiple scales. However, the stride operations in DeepLabv3 can lead to the loss of fine details at object boundaries. To overcome this limitation, DeepLabv3+ [49] introduces a concise yet effective decoder into DeepLabv3, dramatically improving semantic segmentation results, particularly along label boundaries.

Unlike prior works [42, 52] that focus on the recovery of high-resolution feature maps from lower resolutions, the high-resolution network (HRNet) [53] preserves high-resolution representations throughout the entire feature extraction and fusion process. This design achieves more accurate predictions by employing progressive and repetitive multi-scale feature fusions through parallel multi-resolution sub-networks. PointRend [54] introduces a novel point-based rendering technique within a neural network module for image segmentation. By making predictions at adaptively selected locations determined through an iterative subdivision algorithm, PointRend enables precise and flexible segmentation, applicable to both instance and semantic segmentation tasks.

Attention mechanisms have become integral to recent scene parsing networks, significantly improving their ability to focus on relevant features within an image. However, their extensive computational demands have posed significant limitations for deployment. To address this challenge, the asymmetric non-local neural network (ANN) [55] samples only a few representative points from the feature maps, drastically reducing computational complexity. Furthermore, traditional attention mechanisms involve both pairwise and unary terms, which can be difficult to optimize independently. The disentangled non-local network (DNLNet) [56] effectively addresses this challenge by decoupling the interdependence between these two components, enabling more efficient learning and application. Building on prior works, the global context network (GCNet) [57] provides a simplified query-independent formulation, preserving the accuracy of non-local networks while significantly reducing computational overhead.

While attention mechanisms have outperformed traditional approaches, such as ASPP [48], large convolutional kernels, and stacked convolutional layers, in terms of segmentation performance, their substantial GPU memory demands often make them prohibitively expensive. Therefore, several networks have been developed to minimize computational requirements. The interlaced sparse self-attention network (ISANet) [58] factorizes the dense affinity matrix into the product of two sparse matrices, thereby reducing the computational load. Unlike methods that treat all pixels as reconstruction bases [59, 60], the expectation maximization attention network (EMANet) [61] identifies a more compact basis set, significantly decreasing computational complexity. This approach effectively simplifies the creation of large attention maps while also substantially reducing memory consumption. Furthermore, CGNet [62] introduces a parameter-efficient context-guided (CG) block that integrates local features with surrounding context and refines them using global context. Leveraging this module, CGNet delivers competitive performance on the Cityscapes dataset [31] with fewer than 0.5 million parameters.

The vision Transformer (ViT) [63] has been gaining momentum in recent years, particularly for semantic segmentation. The segmentation Transformer (SETR) [64] is the first Transformer-based general-purpose semantic segmentation network. Building on the success of ViT [63], SETR tokenizes images into patches that are processed through Transformer blocks. The encoded features are then gradually upsampled through convolutions to achieve pixel-level classification. SegFormer [34] introduces a multi-scale Transformer encoder for semantic segmentation, which stacks Transformer blocks and inserts convolutional layers between them. Compared to SETR, SegFormer greatly improves segmentation performance, particularly when handling objects of varying sizes. Swin Transformer employs a hierarchical Transformer architecture that computes representations with shifted windows. Inspired by DETR [65], Segmenter develops a mask Transformer decoder that captures global context effectively during both encoding and decoding stages. To improve semantic segmentation at both global and local scales, Twins [66] adopts a dual-branch architecture that captures global contextual information in one branch and focuses on local boundary details in the other. DPT [67] uses ViT as the backbone, assembling tokens at different resolutions from multiple stages of ViT and progressively combining them into full-resolution predictions via a convolutional decoder. ViT processes images at a constant, relatively high resolution, enabling a global receptive field at every stage, which allows the generation of finer-grained and more globally coherent predictions. Similar to the self-attention mechanism used in ViT, object-contextual representation (OCR) [68] characterizes pixels by exploiting the representations of corresponding object classes. While traditional multi-scale context schemes differentiate pixels based on spatial positions, OCR distinguishes between contextual pixels of the same object class and those from different classes. Inspired by Transformer-based architectures, K-Net [69] addresses various image segmentation tasks, including semantic, instance, and panoptic segmentation. This is achieved primarily through a set of learnable kernels, where each kernel generates a mask for either a potential instance or a stuff class.

In contrast to the Transformer-based networks discussed above, MaskFormer [70] introduces a novel paradigm for semantic segmentation that moves beyond traditional per-pixel classification. This architecture performs semantic segmentation by decoding query features into class-specific masks. Specifically, MaskFormer utilizes a multi-scale Transformer decoder that simultaneously generates masks for each class using refined queries, demonstrating superior performance over previous per-pixel classification approaches. Similarly, Mask2Former [71] extends these capabilities to a broader range of image segmentation tasks using a Transformer-based architecture. Built upon MaskFormer [70], it further employs a masked attention mechanism that greatly improves network performance across various segmentation tasks.

3.2 Feature-fusion models

Feature-fusion models effectively leverage heterogeneous features extracted from both RGB images and spatial geometric data, such as depth, surface normal, and transformed disparity, to improve scene parsing performance. FuseNet [27] is among the first to integrate depth information into scene parsing. It employs separate CNN encoders for RGB and depth images and fuses their features via element-wise summation. MFNet [72] achieves a balance between speed and accuracy in driving scene parsing through RGB-thermal data fusion. Similarly, RTFNet [73] utilizes RGB-thermal data as inputs and develops a robust decoder that uses shortcuts to produce clear boundaries while retaining detailed features. While general-purpose feature-fusion models can be applied to road scene parsing, task-specific approaches [11, 26, 28] have been shown to consistently deliver superior performance.

4 Task-specific road scene parsing models

This section provides a brief review of SoTA task-specific road scene parsing models, which can be categorized into 2D, 3D, and hybrid approaches. Their performance is evaluated both quantitatively and qualitatively in Sect. 5.

Early task-specific road scene parsing methods [74–77] predominantly rely on RGB images. Notable advancements in this area include HA-DeepLabv3+ [78] and LFD-RoadSeg [79]. The former introduces an innovative data augmentation strategy based on stereo homography, which generates synthetic images from target images as if viewed from a reference perspective. This method significantly outperforms the baseline DeepLabv3+ [80] and other SoTA stereo vision-based approaches. The recent study [79] introduces LFD-RoadSeg, a two-branch drivable area detection model. The first branch extracts low-level features using the initial stages of ResNet-18 [81], while the second branch enhances contextual recognition by simultaneously downsampling the image and aggregating features. This feature extraction and aggregation strategy enables receptive fields comparable to those of the third stage of ResNet-18, while significantly reducing computational time. A selective fusion module then calculates pixel-wise attention between the low-level representation and contextual features to effectively and efficiently distinguish between road and non-road areas. Nevertheless, these 2D approaches remain highly sensitive to environmental factors such as illumination and weather conditions [26].

With the increasing use of range sensors, particularly LiDARs, 3D approaches [82, 83] have become more robust for road scene parsing. The study [82] presents a deep learning approach for drivable area detection that relies solely on LiDAR data, where unstructured point clouds are transformed into top-view images. These images capture basic statistics, such as mean elevation and density, thereby simplifying the drivable area detection task into a single-scale problem. Subsequently, the study [83] presents a CNN model designed specifically for LiDAR-based semantic segmentation. A computationally efficient hardware architecture is developed and deployed on an FPGA, achieving a processing time of only 17.59 ms per LiDAR scan.

Moreover, LiDAR-camera data fusion approaches have recently become the predominant methods in this domain. A research group has introduced a series of networks [84–86] to solve this problem. For example, the study [84] introduces a two-view fusion-based CNN designed specifically for drivable area detection. This network processes two transformed representations of LiDAR data to deliver pixel-wise drivable area detection results in both LiDAR imagery and camera perspective views in an end-to-end fashion. A mapping layer is incorporated to transfer features from the LiDAR imagery view to the camera perspective view, thereby enhancing the network performance by leveraging the data associations between these two representations. This method optimizes the utilization of LiDAR data, resulting in more accurate and comprehensive drivable area detection results in urban environments. Furthermore, they introduce a drivable area detection approach [85] that integrates LiDAR and camera data within a conditional random field (CRF) framework, combining both range and color information to im-

Table 1 Quantitative comparisons of SoTA road scene parsing networks on the Cityscapes dataset [31]. The symbol ↑ indicates higher values correspond to better performance, while ↓ implies the opposite. 'RGB': RGB images, and 'Normal': surface normal maps

Method	Input data	**IoU** (%) ↑	Fsc (%) ↑	Pre (%) ↑	Rec (%) ↑	mIoU (%) ↑	**Rank** ↓
Mask2Former [71]	RGB	93.84	96.82	97.14	96.51	74.80	5
SegFormer [34]	RGB	93.98	96.90	96.02	97.79	64.51	4
DeepLabv3+ [80]	RGB	93.82	96.81	96.99	96.63	68.66	6
HRNet [87]	RGB	94.06	96.94	96.29	97.59	70.10	3
FuseNet [27]	RGB + Normal	91.60	95.60	96.00	95.30	52.70	9
SNE-RoadSeg [26]	RGB + Normal	93.80	96.80	96.10	97.50	53.40	7
RTFNet [73]	RGB + Normal	94.10	96.90	96.30	97.60	49.60	2
OFF-Net [23]	RGB + Normal	89.60	94.50	93.40	95.70	39.20	10
MFNet [72]	RGB + Normal	92.10	95.90	94.10	97.70	49.30	8
RoadFormer [11]	RGB + Normal	**95.80**	**97.86**	**97.74**	**97.97**	**76.20**	**1**

prove accuracy. In the LiDAR component, a rapid height-difference scanning strategy is applied within the 2D LiDAR range-image domain, enabling precise drivable area detection in the camera image domain through geometric upsampling, which relies on accurate LiDAR-camera calibration. Concurrently, the camera component uses an FCN to process RGB images. The fusion of detailed and binary drivable area detection outputs from both LiDAR and camera data is achieved through a unified CRF framework, effectively optimizing the use of multi-modal data for robust drivable area detection. They further developed CLCFNet [86], a cascaded LiDAR-camera fusion strategy that operates in two modes: a single-modal mode, which uses only LiDAR data, and a multi-modal mode, which combines both LiDAR and camera data to adapt to varying lighting conditions. The network architecture consists of three main components: a LiDAR segmentation module that detects road points from LiDAR data, a sparse-to-dense module that enhances the resolution of LiDAR feature maps for more precise drivable area detection, and a LiDAR-camera fusion module that integrates these high-resolution maps with camera images to provide accurate road estimations. This framework is designed to deliver robust and precise drivable area detection across diverse environmental conditions.

Other representative LiDAR-camera data-fusion approaches in road scene parsing include LidCamNet [88], PLARD [89], PLB-RD [90], and USNet [91]. In 2018, the study [88] introduces LidCamNet, a deep learning-based drivable area detection framework that utilizes both LiDAR point clouds and camera images. First, unstructured and sparse LiDAR point clouds are projected onto the camera image plane and upsampled to generate dense 2D images that capture spatial details. Road detection is then performed using several FCNs, which operate on data from either a single sensor or through one of three fusion strategies: early-fusion, late-fusion, and cross-fusion. In the early and late fusion strategies, multi-modal information is fused at specific depth levels within the network.

Table 2 Quantitative comparisons of SoTA road scene parsing networks on the KITTI Semantics dataset [39]. The symbol ↑ indicates higher values correspond to better performance, while ↓ implies the opposite

Method	Fsc (%) ↑	**IoU** (%) ↑	Acc (%) ↑
NIM-RTFNet [92]	92.59	85.95	96.61
OFF-Net [23]	93.82	86.79	97.08
SNE-RoadSeg [26]	94.85	88.02	97.33
SNE-RoadSeg+ [28]	95.11	89.07	97.59
RoadFormer [11]	95.36	90.18	97.83
SNE-RoadSegV2 [6]	**96.60**	**91.75**	**98.44**

The cross-fusion FCN, on the other hand, identifies optimal points for integrating data across the LiDAR and camera branches via trainable cross-connections. This improves the system's ability to capture and leverage complex spatial relationships, thereby improving drivable area detection accuracy. In 2019, the study [89] introduces PLARD, a drivable area detection strategy, which enhances image-based road detection by integrating LiDAR data. PLARD employs two primary modules: 1) data space adaptation, where LiDAR data is transformed to align with the visual data space through altitude difference-based transformations to match the perspective view, and 2) feature space adaptation, which integrates LiDAR features with visual features through a cascaded fusion structure to optimize detection performance. In [90], LRDNet+ is introduced to address the challenge of integrating LiDAR and visual features, which exist in different spaces, by learning transformation and fusion operations that enhance visual features with LiDAR data. Subsequently, the study [96] introduces an RGB-LiDAR drivable area detection method, referred to as PLB-RD. By simulating LiDAR through depth estimation, the approach uses a feature fusion network that integrates RGB images with derived depth information to enhance drivable area detection accuracy. A strategy is also developed to optimize information flow pathways. Additionally, a modality distillation strategy is employed to minimize computational demands during model inference, which eliminates the need

Table 3 Quantitative comparisons of SoTA general-purpose semantic segmentation networks on the KITTI Road dataset [29]. The symbol ↑ indicates higher values correspond to better performance, while ↓ implies the opposite

Method	**IoU** (%) ↑	Acc (%) ↑	Fscore (%) ↑	Pre (%) ↑	Rec (%) ↑	**Rank** ↓
ANN [55]	94.94	98.55	97.41	96.28	98.55	8
CGNet [62]	92.26	96.93	95.98	95.04	96.93	17
DNLNet [93]	95.16	98.64	97.52	96.42	98.64	3
DPT [67]	94.55	98.16	97.20	96.25	98.16	13
EMANet [94]	94.87	98.64	97.37	96.13	98.64	11
Fast-SCNN [44]	90.24	95.17	94.87	94.57	95.17	19
FastFCN [43]	94.95	98.42	97.41	96.42	98.42	7
GCNet [57]	94.92	98.43	97.39	96.38	98.43	10
ISANet [58]	94.93	98.42	97.40	96.40	98.42	9
K-Net [69]	95.32	98.72	97.61	96.52	98.72	2
Mask2Former [71]	**95.34**	**98.90**	**97.62**	96.37	**98.90**	**1**
MaskFormer [70]	95.12	98.63	97.50	96.39	98.63	4
OCRNet [68]	94.83	98.25	97.35	96.46	98.25	12
PIDNet [45]	95.08	98.39	97.48	**96.58**	98.39	5
PointRend [54]	94.99	98.49	97.43	96.40	98.49	6
SegFormer [34]	88.67	97.06	94.00	91.12	97.06	20
Segmenter [95]	92.73	96.92	96.23	95.55	96.92	16
SETR [64]	92.19	97.47	95.94	94.45	97.47	18

Table 4 Quantitative comparisons of SoTA road scene parsing networks on the KITTI Road dataset [29]. These results are publicly available at cvlibs.net/datasets/kitti/eval_road.php. The symbol ↑ indicates that higher values correspond to better performance, while ↓ implies the opposite. 'RGB': RGB images, 'Disp': disparity images, 'Depth': depth images, 'PC': LiDAR point clouds, and 'Normal': surface normal maps

Method	Input data	**MaxF** (%) ↑	AP (%) ↑	Pre (%) ↑	Rec (%) ↑	FPR (%) ↓	FNR (%) ↓	**Rank** ↓
LFD-RoadSeg [79]	RGB	95.21	93.71	95.35	95.08	2.56	4.92	26
HA-DeepLabv3+ [78]	RGB + Disp	94.83	93.24	94.77	94.89	2.88	5.11	32
DFM-RTFNet [7]	RGB + Disp	96.78	94.05	96.62	96.93	1.87	3.07	11
USNet [91]	RGB + Depth	96.89	93.25	96.51	97.27	1.94	2.73	9
LRDNet+ [90]	RGB + PC	96.95	92.22	96.88	97.02	1.72	2.98	8
PLB-RD [96]	RGB + PC	97.42	**94.09**	97.30	97.54	1.49	2.46	5
PLARD [89]	RGB + PC	97.03	94.03	97.19	96.88	1.54	3.12	7
LidCamNet [88]	RGB + PC	96.03	93.93	96.23	95.83	2.07	4.17	17
CLCFNet [86]	RGB + PC	96.38	90.85	96.38	96.39	1.99	3.61	15
TVFNet [84]	RGB + PC	95.34	90.26	95.73	94.94	2.33	5.06	25
ChipNet [83]	RGB + PC	94.05	88.29	93.57	94.53	3.58	5.47	37
LoDNN [79]	RGB + PC	94.07	92.03	92.81	95.37	4.07	4.63	36
NIM-RTFNet [92]	RGB + Normal	96.02	94.01	96.43	95.62	1.95	4.38	18
SNE-RoadSeg [26]	RGB + Normal	96.75	94.07	96.90	96.61	1.70	3.39	12
SNE-RoadSeg+ [28]	RGB + Normal	97.50	93.98	97.41	97.58	1.43	2.42	4
RoadFormer [11]	RGB + Normal	97.50	93.85	97.16	**97.84**	1.57	**2.16**	3
SNE-RoadSegV2 [6]	RGB + Normal	97.55	93.98	**97.57**	97.53	**1.34**	2.47	2
RoadFormer+ [97]	RGB + Normal	**97.56**	93.74	97.43	97.69	1.42	2.31	**1**

for depth estimation networks at this stage. USNet [91] effectively balances speed and accuracy in drivable area detection by leveraging both RGB and depth data without relying on traditional cross-modal feature fusion. Instead, it employs two lightweight sub-networks to process RGB and depth data independently, ensuring real-time performance. A multi-scale evidence collection module gathers evidence from each modality across different scales to improve pixel-level classification. An uncertainty-aware fusion module then uses the perceived uncertainty of each modality to guide the integration of sub-network outputs, thereby enhancing segmentation accuracy.

Additionally, inspired by FuseNet [27], SoTA approaches generally adopt duplex-encoder architectures [11, 26, 92], where each encoder extracts hierarchical features from a specific data source or modality. The extracted heterogeneous features are subsequently fused, enabling the network to gain a more comprehensive understanding of the environment [89]. For example, NIM-RTFNet [92], SNE-RoadSeg [26], SNE-RoadSeg+ [28], and SNE-RoadSegV2

[6] incorporate surface normal information into drivable area detection. This series employs densely-connected skip connections to enhance feature extraction in the decoder, thereby achieving SoTA performance compared to other approaches. Drawing on the success of single-modal Transformers, OFF-Net [23] is the first attempt to apply a Transformer architecture for feature-fusion road scene parsing. It utilizes a SegFormer [34] encoder to generate RGB and surface normal features, outperforming SoTA CNNs in off-road drivable area detection. Expanding upon these foundational previous studies, RoadFormer [11] also adopts the feature-fusion paradigm but distinguishes itself by employing a novel Transformer architecture for road scene parsing. RoadFormer incorporates a unique feature synergy block, which significantly enhances the overall performance across multiple road scene parsing datasets, outperforming all other feature-fusion networks.

As for the input data, the most commonly used spatial geometric information includes depth/disparity maps [91, 98], LiDAR point clouds [89, 90], and surface normal maps [11, 26, 28]. Extensive experiments conducted in previous studies [7, 11, 26, 28] have consistently demonstrated that surface normal maps and transformed disparity maps provide the most informative spatial geometric features for road scene parsing, primarily due to their ability to represent planar characteristics.

Table 5 Quantitative comparisons of SoTA road scene parsing networks on the ORFD [23] dataset. The results are sourced from the original paper [23] and re-evaluated in the study [11]. The symbol ↑ indicates that higher values correspond to better performance

Method	**IoU** (%) ↑	Fsc (%) ↑	Pre (%) ↑	Rec (%) ↑
Published				
FuseNet [27]	66.00	79.50	74.50	85.20
SNE-RoadSeg [26]	81.20	89.60	86.70	92.70
OFF-Net [23]	82.30	90.30	86.60	94.30
Re-implemented				
FuseNet [27]	59.00	74.20	59.30	**99.10**
SNE-RoadSeg [26]	79.50	88.60	90.30	86.90
RTFNet [92]	90.70	95.10	93.80	96.50
OFF-Net [23]	81.80	90.00	84.20	96.70
MFNet [72]	81.70	89.90	89.60	90.30
RoadFormer [11]	**92.51**	**96.11**	**95.08**	97.17

5 Comprehensive comparisons

The quantitative comparisons of SoTA road scene parsing networks are given in Tables 1, 2, 3, 4, 5, and 6. Their corresponding qualitative results are illustrated in Figs. 2, 3, 4, 5, 6, and 7, respectively.

As anticipated, fusing heterogeneous features extracted from multiple data sources yields superior performance compared to using features from a single data type. For instance, RoadFormer significantly outperforms its baseline network, Mask2Former, across various datasets, with particularly notable improvements on the SYN-UDTIRI dataset.

Second, it has been observed that task-specific networks generally outperform universal semantic segmentation networks trained for binary or ternary pixel classifications. This improvement can be attributed to two key factors: data type and architecture. Task-specific networks often leverage informative spatial geometric data, such as surface normals and transformed disparity maps, to enhance road scene parsing performance. Additionally, architectures incorporating robust and effective feature fusion modules enable a more comprehensive understanding of the road scene, thereby significantly improving segmentation accuracy.

Finally, while road surface detection can now be considered a well-solved problem due to significant improvements in accuracy, road defect detection remains a challenging area requiring further research. Current results are still below satisfactory standards, with state-of-the-art

Table 6 Quantitative comparisons of road defect detection using SoTA road scene parsing networks on the SYN-UDTIRI dataset [11]. The symbol ↑ indicates that higher values correspond to better performance, while ↓ implies the opposite. 'RGB': RGB images, and 'Normal': surface normal maps

Method	Input data	**IoU** (%) ↑	Fsc (%) ↑	Pre (%) ↑	Rec (%) ↑	**Rank** ↓
Mask2Former [71]	RGB	46.91	63.87	73.59	56.41	7
SegFormer [34]	RGB	36.34	53.31	57.23	49.89	8
DeepLabv3+ [80]	RGB	34.76	51.58	62.54	43.90	10
HRNet [87]	RGB	35.47	52.37	69.09	42.16	9
CGNet [62]	RGB	19.36	32.44	28.10	38.37	12
K-Net [69]	RGB	33.15	49.79	63.01	41.15	11
FuseNet [27]	RGB + Normal	70.70	82.90	72.10	**97.50**	6
RTFNet [73]	RGB + Normal	90.50	95.00	95.50	94.50	3
OFF-Net [23]	RGB + Normal	83.80	91.20	91.90	90.50	5
MFNet [72]	RGB + Normal	87.70	93.50	96.20	90.90	4
SNE-RoadSeg [26]	RGB + Normal	92.10	95.90	**96.70**	95.10	2
RoadFormer [11]	RGB + Normal	**93.51**	**96.65**	96.61	96.69	**1**

Figure 2 Qualitative comparisons of SoTA road scene parsing networks on the Cityscapes dataset [31], with road classifications shown in purple

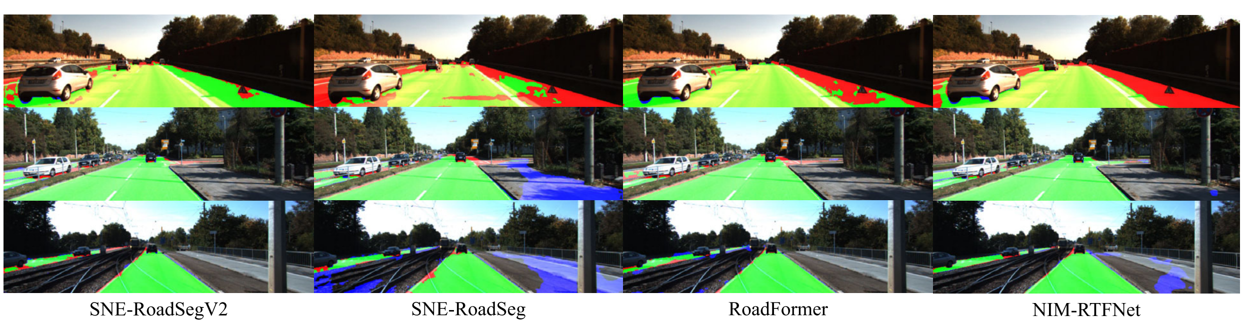

Figure 3 Qualitative comparisons of SoTA road scene parsing networks on the KITTI Semantics dataset [39], with true-positive, false-positive, and false-negative classifications shown in green, blue, and red, respectively

single-modal networks achieving IoU scores of only 19% to 47% on the SYN-UDTIRI dataset.

6 Discussion

Despite their compelling results, existing road scene parsing approaches still face several key limitations.

First, a significant limitation of feature-fusion networks is their reliance on spatial geometric information, such as 3D point clouds from LiDAR, which restricts their applicability in environments lacking LiDAR sensors. Furthermore, inaccuracies in the data, such as variations in camera-LiDAR calibration, can degrade the fusion of heterogeneous features, ultimately reducing the overall performance of road scene parsing. As a result, online, target-free LiDAR-camera extrinsic calibration is both necessary and critical [99]. While stereo cameras provide a practical and cost-effective alternative to LiDAR for obtaining depth information, incorporating a separate stereo matching network increases computational demands, making

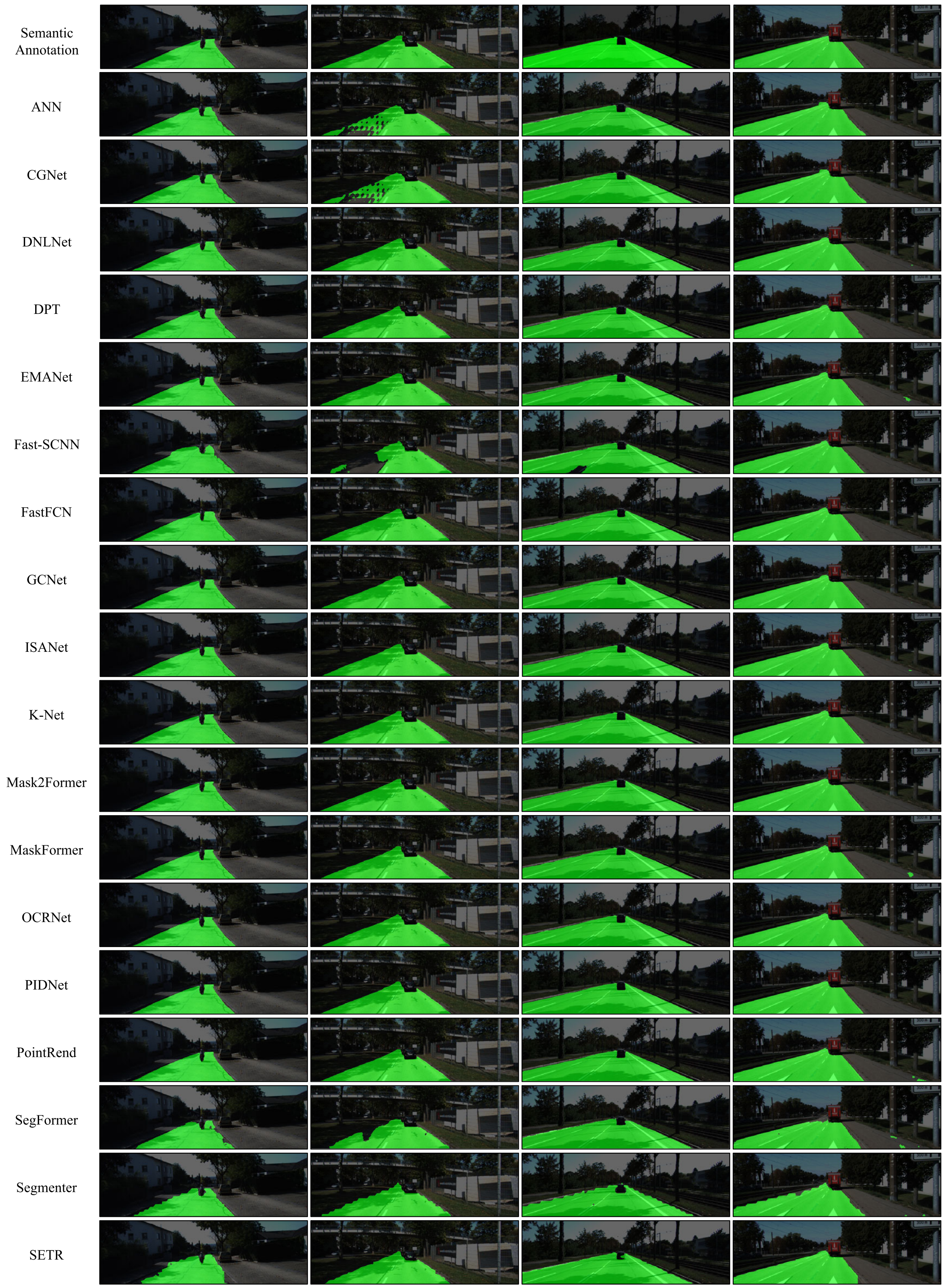

Figure 4 Qualitative comparisons of SoTA general-purpose semantic segmentation networks on the KITTI Road dataset [29], with road classifications shown in green

Figure 5 Qualitative comparisons of SoTA road scene parsing networks on the KITTI Road dataset [29], with true-positive, false-positive, and false-negative classifications shown in green, blue, and red, respectively

real-time processing challenging [100]. Additionally, the requirement for spatial geometric information continues to hinder deployment in sensor-constrained environments. Therefore, there is an urgent need to improve the

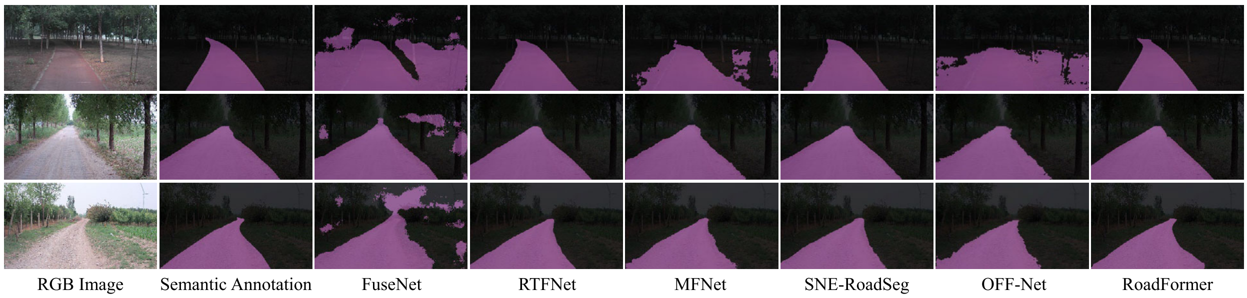

Figure 6 Qualitative comparisons of SoTA road scene parsing networks on the ORFD [23] dataset, with road classifications shown in purple

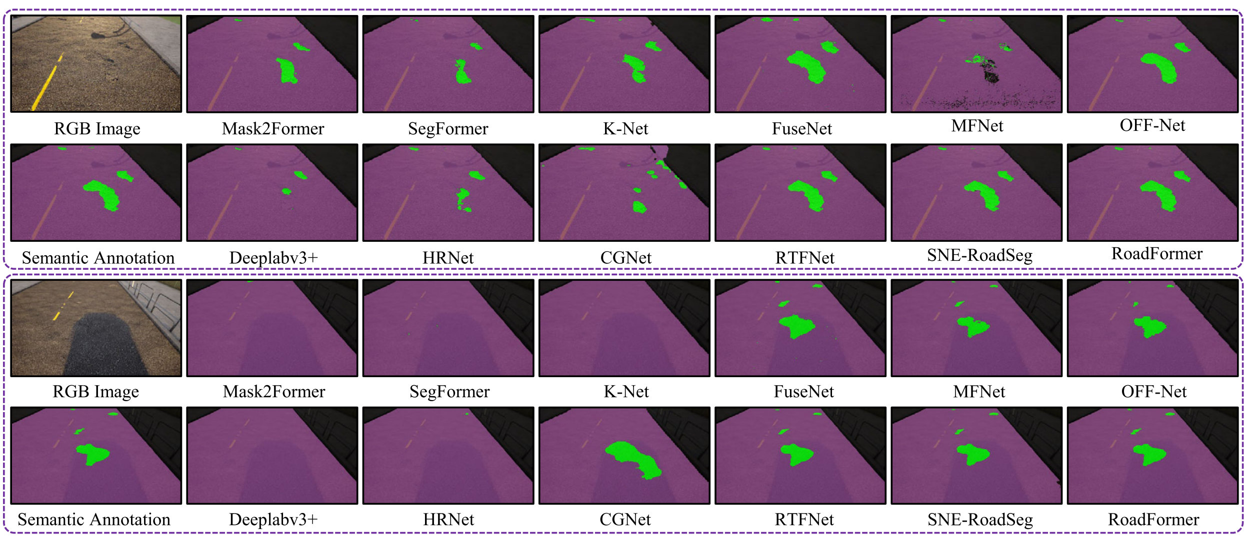

Figure 7 Qualitative comparisons of SoTA road scene parsing networks on the SYN-UDTIRI dataset [11], with drivable areas shown in purple and road defects shown in green

robustness of single-modal methods and develop effective cross-modal knowledge distillation techniques (from RGB+X to RGB) to address these constraints.

Second, feature-fusion networks are typically computationally intensive, posing challenges for deployment on edge-computing devices with limited resources. Thus, designing lightweight network architectures, such as those based on MobileNet or EfficientNet, is crucial for platforms requiring real-time operation. Additionally, networks optimized with TensorRT or similar technologies are essential in robotics and autonomous driving, where rapid data processing is critical.

Third, the performance of existing networks on public datasets has nearly reached its ceiling, with recent architectural innovations yielding only marginal improvements in accuracy. While several synthetic datasets [11] have been developed to simulate diverse lighting conditions and adverse weather scenarios, they remain limited in capturing the complexity and nuanced characteristics of real-world environments, particularly regarding texture fidelity and scene authenticity. This limitation is evident in experimental results from [11, 97], where feature-fusion networks show performance saturation on synthetic datasets but experience significant degradation when tested on more challenging real-world data. Moreover, networks pre-trained on synthetic datasets exhibit substantial performance variability when deployed in real-world scenarios, highlighting the inherent shortcomings of current synthetic datasets. These observations underscore the urgent need for new synthetic or real-world datasets that incorporate multi-modal data and detailed annotations for various road elements.

Additionally, the establishment of new benchmarks beyond KITTI and Cityscapes is crucial to guide researchers in developing innovative networks specifically for this task. While supervised learning currently dominates the field of road scene parsing, there is significant untapped potential in adapting and systematically evaluating semi-supervised and self-supervised learning paradigms for this domain. Recent studies have demonstrated promising results, such as unsupervised frameworks that utilize generative adversarial networks and multi-scale masking techniques for annotation-free crack detection. These methods could be effectively extended to broader road defect detection tasks. Exploring and benchmarking such alternative learning paradigms would not only expand the methodological toolkit available to researchers but also address fundamental limitations of purely supervised approaches, particularly in scenarios where labeled data is scarce or expensive to obtain.

7 Conclusion

This article provided a brief review of state-of-the-art CNNs and Transformers developed for general-purpose semantic segmentation and task-specific road scene parsing, gave comprehensive comparisons of their performance across five public datasets, and discussed existing challenges and future trends in this research field. We hope this review will serve as a valuable resource for readers and researchers, guiding the development of next-generation models for road scene parsing.

Acknowledgements
We gratefully acknowledge the financial support of Xiaomi Corporation.

Author contributions
All authors contributed to drafting the manuscript and discussing improvements. RF, JL, JLi, and JW were involved in conducting the experiments and performing the experimental analyses.

Funding
This research was supported by the National Natural Science Foundation of China under Grant Nos. 62473288, 62403361, and 62233013, the Fundamental Research Funds for the Central Universities, and Xiaomi Young Talents Program. It should be clarified that the collaboration among the authors is limited solely to this work and does not extend to any other projects.

Data availability
The data utilized in this study are publicly accessible.

Declarations

Competing interests
The authors declare that they have no competing interests.

Author details
[1]College of Electronics & Information Engineering, Shanghai Institute of Intelligent Science and Technology, Shanghai Research Institute for Intelligent Autonomous Systems, the State Key Laboratory of Intelligent Autonomous Systems, and the Frontiers Science Center for Intelligent Autonomous Systems, Tongji University, 4800 Caoan Rd, Shanghai, 201804, China. [2]College of Electronics & Information Engineering, Tongji University, 4800 Caoan Rd, Shanghai, 201804, China. [3]School of Microelectronics, Shanghai University, 20 Chengzhong Rd, Shanghai, 201800, China. [4]School of Mechanical Engineering, Beijing Institute of Technology, 5 Zhongguancun South Street, Beijing, 100081, China. [5]Department of Electrical and Electronic Engineering, University of Bristol, Woodland Road, Bristol, BS8 1UB, England, UK. [6]Faculty of Radio Engineering and Telecommunications, Ryazan State Radio Engineering University, Ulitsa Gagarina, 59/1, Ryazan, 390005, Ryazan Oblast, Russian Federation. [7]College of Intelligence Science and Technology, National University of Defense Technology, 109 Deya Road, Changsha, 410073, China. [8]Microwave Photonics Department, Telecommunications Center, Moscow Institute of Physics and Technology, Institutsky Lane, 9, Dolgoprudny, 141701, Moscow Region, Russian Federation. [9]Multimedia Technology and Telecom Department, Telecommunications Center, Moscow Institute of Physics and Technology, Institutsky Lane, 9, Dolgoprudny, 141701, Moscow Region, Russian Federation.

Received: 10 November 2024 Revised: 21 January 2025
Accepted: 27 February 2025 Published online: 13 March 2025

References

1. M. Weber, et al., Approach for improved development of advanced driver assistance systems for future smart mobility concepts. Auton. Intell. Syst. **3**(1), 2 (2023)
2. K. Yuan, et al., Human feedback enhanced autonomous intelligent systems: a perspective from intelligent driving. Auton. Intell. Syst. **4**(1), 1–10 (2024)
3. C.W. Liu, et al., These maps are made by propagation: adapting deep stereo networks to road scenarios with decisive disparity diffusion. IEEE Trans. Image Process. **34**, 1516–1528 (2025)
4. M.O. Macaulay, M. Shafiee, Machine learning techniques for robotic and autonomous inspection of mechanical systems and civil infrastructure. Auton. Intell. Syst. **2**(1), 8 (2022)
5. M.A. Goodale, Lessons from human vision for robotic design. Auton. Intell. Syst. **1**(1), 2 (2021)
6. Y. Feng, et al., SNE-RoadSegV2: advancing Heterogeneous Feature Fusion and Fallibility Awareness for Freespace Detection. IEEE Trans. Instrum. Meas. (2025). https://doi.org/10.1109/TIM.2025.3545498
7. H. Wang, et al., Dynamic fusion module evolves drivable area and road anomaly detection: a benchmark and algorithms. IEEE Trans. Cybern. **52**(10), 10750–10760 (2021)
8. M. Rubagotti, et al., Perceived safety in physical human–robot interaction—a survey. Robot. Auton. Syst. **151**, 104047 (2022)
9. X. Zhang, et al., An intelligent surface roughness prediction method based on automatic feature extraction and adaptive data fusion. Auton. Intell. Syst. **4**(1), 1–17 (2024)
10. S. KC, Enhanced pothole detection system using yolox algorithm. Auton. Intell. Syst. **2**(1), 22 (2022)
11. J. Li, et al., RoadFormer: duplex transformer for RGB-normal semantic road scene parsing. IEEE Trans. Intell. Veh. **9**(7), 5163–5172 (2024)
12. S. Guo, et al., UDTIRI: an online open-source intelligent road inspection benchmark suite. IEEE Trans. Intell. Transp. Syst. **25**(8), 9920–9931 (2024)
13. R. Fan, et al., Pothole detection based on disparity transformation and road surface modeling. IEEE Trans. Image Process. **29**, 897–908 (2020)
14. M. Fan, et al., Rethinking BiSeNet for real-time semantic segmentation, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2021), pp. 9716–9725
15. A. Wedel, et al., B-spline modeling of road surfaces for freespace estimation, in *2008 IEEE Intelligent Vehicles Symposium (IV)* (IEEE, 2008), pp. 828–833
16. A. Wedel, et al., B-spline modeling of road surfaces with an application to free-space estimation. IEEE Trans. Intell. Transp. Syst. **10**(4), 572–583 (2009)
17. A. Rasheed, et al., Stabilization of 3D pavement images for pothole metrology using the Kalman filter, in *IEEE International Conference on Intelligent Transportation Systems (ITSC)* (IEEE, 2015), pp. 2671–2676
18. G.L. Oliveira, et al., Efficient deep models for monocular road segmentation, in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (IEEE, 2016), pp. 4885–4891
19. L. Chen, et al., LiDAR-histogram for fast road and obstacle detection, in *2017 IEEE International Conference on Robotics and Automation (ICRA)* (IEEE, 2017), pp. 1343–1348
20. R. Fan, et al., Road surface 3D reconstruction based on dense subpixel disparity map estimation. IEEE Trans. Image Process. **27**(6), 3025–3035 (2018)
21. A. Dhiman, R. Klette, Pothole detection using computer vision and learning. IEEE Trans. Intell. Transp. Syst. **21**(8), 3536–3550 (2019)
22. R. Fan, et al., Graph attention layer evolves semantic segmentation for road pothole detection: a benchmark and algorithms. IEEE Trans. Image Process. **30**, 8144–8154 (2021)

23. C. Min, et al., ORFD: a dataset and benchmark for off-road freespace detection, in *IEEE International Conference on Robotics and Automation (ICRA)* (IEEE, 2022), pp. 2532–2538
24. B. Tian, et al., Unsupervised road anomaly detection with language anchors, in *2023 IEEE International Conference on Robotics and Automation (ICRA)* (IEEE, 2023), pp. 7778–7785
25. C. Lu, et al., Monocular semantic occupancy grid mapping with convolutional variational encoder–decoder networks. IEEE Robot. Autom. Lett. **4**(2), 445–452 (2019)
26. R. Fan, et al., SNE-RoadSeg: incorporating surface normal information into semantic segmentation for accurate freespace detection, in *Proceedings of the European Conference on Computer Vision (ECCV)* (Springer, Berlin, 2020), pp. 340–356
27. C. Hazirbas, et al., FuseNet: incorporating depth into semantic segmentation via fusion-based CNN architecture, in *Proceedings of the Asian Conference on Computer Vision (ACCV)* (Springer, Berlin, 2017), pp. 213–228
28. H. Wang, et al., SNE-RoadSeg+: rethinking depth-normal translation and deep supervision for freespace detection, in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (IEEE, 2021), pp. 1140–1145
29. J. Fritsch, et al., A new performance measure and evaluation benchmark for road detection algorithms, in *IEEE International Conference on Intelligent Transportation Systems (ITSC)* (IEEE, 2013), pp. 1693–1700
30. Y. Cabon, et al., Virtual KITTI 2. Comput. Res. Repos. (CoRR) (2020). arXiv:2001.10773
31. M. Cordts, et al., The cityscapes dataset for semantic urban scene understanding, in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2016), pp. 3213–3223
32. K. Han, et al., A survey on vision transformer. IEEE Trans. Pattern Anal. Mach. Intell. **45**(1), 87–110 (2022)
33. Z. Liu, et al., Swin transformer: hierarchical vision transformer using shifted windows, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (2021), pp. 10012–10022
34. E. Xie, et al., SegFormer: simple and efficient design for semantic segmentation with transformers. Adv. Neural Inf. Process. Syst. **34**, 12077–12090 (2021)
35. K. Li, et al., UniFormer: unifying convolution and self-attention for visual recognition. IEEE Trans. Pattern Anal. Mach. Intell. **45**(10), 12581–12600 (2023)
36. L. Lipson, et al., RAFT-stereo: multilevel recurrent field transforms for stereo matching, in *International Conference on 3D Vision (3DV)* (IEEE, 2021), pp. 218–227
37. M. Menze, A. Geiger, Object scene flow for autonomous vehicles, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2015), pp. 3061–3070
38. R. Fan, et al., Three-filters-to-normal: an accurate and ultrafast surface normal estimator. IEEE Robot. Autom. Lett. **6**(3), 5405–5412 (2021)
39. A. Geiger, et al., Are we ready for autonomous driving? The KITTI vision benchmark suite, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, 2012), pp. 3354–3361
40. J. Li, et al., Practical stereo matching via cascaded recurrent network with adaptive correlation, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2022), pp. 16263–16272
41. A. Dosovitskiy, et al., CARLA: an open urban driving simulator, in *Conference on Robot Learning (CoRL)* (2017), pp. 1–16. PMLR
42. J. Long, et al., Fully convolutional networks for semantic segmentation, in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2015), pp. 3431–3440
43. H. Wu, et al., FastFCN: Rethinking dilated convolution in the backbone for semantic segmentation. Comput. Res. Repos. (CoRR) (2019). arXiv:1903.11816
44. R.P. Poudel, et al., Fast-SCNN: fast semantic segmentation network, in *The British Machine Vision Conference (BMVC)* (2019)
45. J. Xu, et al., PIDNet: a real-time semantic segmentation network inspired by PID controllers, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2023), pp. 19529–19539
46. A. Kirillov, et al., Panoptic feature pyramid networks, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2019), pp. 6399–6408
47. T. Xiao, et al., Unified perceptual parsing for scene understanding, in *Proceedings of the European Conference on Computer Vision (ECCV)* (2018), pp. 418–434
48. L.-C. Florian, et al., Rethinking atrous convolution for semantic image segmentation, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2017)
49. L.-C. Chen, et al., Encoder-decoder with atrous separable convolution for semantic image segmentation, in *Proceedings of the European Conference on Computer Vision (ECCV)* (2018), pp. 801–818
50. J. He, et al., Dynamic multi-scale filters for semantic segmentation, in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (2019), pp. 3562–3572
51. H. Zhao, et al., Pyramid scene parsing network, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017), pp. 2881–2890
52. O. Ronneberger, et al., U-net: convolutional networks for biomedical image segmentation, in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)* (Springer, Berlin, 2015), pp. 234–241
53. K. Sun, et al., Deep high-resolution representation learning for human pose estimation, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2019), pp. 5693–5703
54. A. Kirillov, et al., PointRend: image segmentation as rendering, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2020), pp. 9799–9808
55. Z. Zhu, et al., Asymmetric non-local neural networks for semantic segmentation, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (2019), pp. 593–602
56. M. Yin, et al., Disentangled non-local neural networks, in *Proceedings of the European Conference on Computer Vision (ECCV)* (Springer, Berlin, 2020), pp. 191–207
57. Y. Cao, et al., GCNet: non-local networks meet squeeze-excitation networks and beyond, in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)* (2019)
58. L. Huang, et al., Interlaced sparse self-attention for semantic segmentation. Comput. Res. Repos. (CoRR) (2019). arXiv:1907.12273
59. X. Wang, et al., Non-local neural networks, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2018), pp. 7794–7803
60. H. Zhao, et al., PSANet: point-wise spatial attention network for scene sarsing, in *Proceedings of the European Conference on Computer Vision (ECCV)* (2018), pp. 267–283
61. X. Li, et al., Expectation-maximization attention networks for semantic segmentation, in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (2019), pp. 9167–9176
62. T. Wu, et al., CGNet: a light-weight context guided network for semantic segmentation. IEEE Trans. Image Process. **30**, 1169–1179 (2020)
63. A. Dosovitskiy, et al., An image is worth 16×16 words: transformers for image recognition at scale, in *International Conference on Learning Representations (ICLR)* (2020)
64. S. Zheng, et al., Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2021), pp. 6881–6890
65. N. Carion, et al., End-to-end object detection with transformers, in *Proceedings of the European Conference on Computer Vision (ECCV)* (Springer, Berlin, 2020), pp. 213–229
66. X. Chu, et al., Twins: revisiting the design of spatial attention in vision transformers. Adv. Neural Inf. Process. Syst. **34**, 9355–9366 (2021)
67. R. Ranftl, et al., Vision transformers for dense prediction, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (2021), pp. 12179–12188
68. Y. Yuan, et al., Object-contextual representations for semantic segmentation, in *Proceedings of the European Conference on Computer Vision (ECCV)* (Springer, Berlin, 2020), pp. 173–190
69. W. Zhang, et al., K-Net: towards unified image segmentation. Adv. Neural Inf. Process. Syst. **34**, 10326–10338 (2021)
70. B. Cheng, et al., Per-pixel classification is not all you need for semantic segmentation. Adv. Neural Inf. Process. Syst. **34**, 17864–17875 (2021)
71. B. Cheng, et al., Masked-attention mask transformer for universal image segmentation, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2022), pp. 1290–1299
72. Q. Ha, et al., MFNet: towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes, in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (IEEE, 2017), pp. 5108–5115

73. Y. Sun, et al., RTFNet: RGB-thermal fusion network for semantic segmentation of urban scenes. IEEE Robot. Autom. Lett. **4**(3), 2576–2583 (2019). https://doi.org/10.1109/LRA.2019.2904733
74. J.M. Alvarez, et al., Road scene segmentation from a single image, in *Proceedings of the European Conference on Computer Vision (ECCV)* (Springer, Berlin, 2012), pp. 376–389
75. L. Xiao, et al., Monocular road detection using structured random forest. Int. J. Adv. Robot. Syst. **13**(3), 101 (2016)
76. C.-A. Brust, et al., Convolutional patch networks with spatial prior for road detection and urban scene understanding, in *International Conference on Computer Vision Theory and Applications (VISAPP)* (2015)
77. D. Levi, et al., StixelNet: a deep convolutional network for obstacle detection and road segmentation, in *Proceedings of the British Machine Vision Conference (BMVC)*, vol. 1 (2015), p. 4
78. R. Fan, et al., Learning collision-free space detection from stereo images: homography matrix brings better data augmentation. IEEE/ASME Trans. Mechatron. **27**(1), 225–233 (2021)
79. H. Zhou, et al., Exploiting low-level representations for ultra-fast road segmentation. IEEE Trans. Intell. Transp. Syst. **25**(8), 9909–9919 (2024)
80. L.-C. Chen, et al., Encoder-decoder with atrous separable convolution for semantic image segmentation, in *Proceedings of the European Conference on Computer Vision (ECCV)* (2018), pp. 801–818
81. K. He, et al., Deep residual learning for image recognition, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016), pp. 770–778
82. L. Caltagirone, et al., Fast LIDAR-based road detection using fully convolutional neural networks, in *IEEE Intelligent Vehicles Symposium (IV)* (IEEE, 2017), pp. 1019–1024
83. Y. Lyu, et al., ChipNet: real-time LiDAR processing for drivable region segmentation on an FPGA. IEEE Trans. Circuits Syst. **66**(5), 1769–1779 (2018)
84. S. Gu, et al., Two-view fusion based convolutional neural network for urban road detection, in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (IEEE, 2019), pp. 6144–6149
85. S. Gu, et al., Road detection through CRF based LiDAR-camera fusion, in *2019 International Conference on Robotics and Automation (ICRA)* (IEEE, 2019), pp. 3832–3838
86. S. Gu, et al., A cascaded LiDAR-camera fusion network for road detection, in *IEEE International Conference on Robotics and Automation (ICRA)* (IEEE, 2021), pp. 13308–13314
87. J. Wang, et al., Deep high-resolution representation learning for visual recognition. IEEE Trans. Pattern Anal. Mach. Intell. **43**(10), 3349–3364 (2020)
88. L. Caltagirone, et al., LiDAR–camera fusion for road detection using fully convolutional neural networks. Robot. Auton. Syst. **111**, 125–131 (2019)
89. Z. Chen, et al., Progressive LiDAR adaptation for road detection. IEEE/CAA J. Autom. Sin. **6**(3), 693–702 (2019)
90. A.A. Khan, et al., LRDNet: Lightweight LiDAR Aided Cascaded Feature Pools for Free Road Space Detection. IEEE Trans. Multimed., 1–13 (2022). https://doi.org/10.1109/TMM.2022.3230330
91. Y. Chang, et al., Fast road segmentation via uncertainty-aware symmetric network, in *IEEE International Conference on Robotics and Automation (ICRA)* (IEEE, 2022), pp. 11124–11130
92. H. Wang, et al., Applying surface normal information in drivable area and road anomaly detection for ground mobile robots, in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (IEEE, 2020), pp. 2706–2711
93. M. Yin, et al., Disentangled non-local neural networks, in *Proceedings of the European Conference on Computer Vision (ECCV)* (Springer, Berlin, 2020), pp. 191–207
94. X. Li, et al., Expectation-maximization attention networks for semantic segmentation, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (2019), pp. 9167–9176
95. R. Strudel, et al., Segmenter: transformer for semantic segmentation, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (2021), pp. 7262–7272
96. L. Sun, et al., Pseudo-LiDAR-based road detection. IEEE Trans. Circuits Syst. Video Technol. **32**(8), 5386–5398 (2022)
97. J. Huang, et al., RoadFormer+: delivering RGB-X scene parsing through scale-aware information decoupling and advanced heterogeneous feature fusion. IEEE Trans. Intell. Veh. (2024). https://doi.org/10.1109/TIV.2024.3448251
98. J.-Y. Sun, et al., Reverse and boundary attention network for road segmentation, in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)* (2019), pp. 876–885
99. Z. Huang, et al., Online, target-free LiDAR-camera extrinsic calibration via cross-modal mask matching. IEEE Trans. Intell. Veh. (2024). https://doi.org/10.1109/TIV.2024.3456299
100. Z. Wu, et al., S^3M-Net: Joint learning of semantic segmentation and stereo matching for autonomous driving. IEEE Trans. Intell. Veh. **9**(2), 3940–3951 (2024)

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.