This article has been accepted for publication in IEEE Transactions on Intelligent Vehicles. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIV.2024.3388726

RoadFormer: Duplex Transformer for RGB-Normal Semantic Road Scene Parsing

Jiahang Li, *Graduate Student Member, IEEE*, Yikang Zhang, *Graduate Student Member, IEEE*, Peng Yun, Guangliang Zhou, Qijun Chen *Senior Member, IEEE*, Rui Fan, *Senior Member, IEEE*

***Abstract*—The recent advancements in deep convolutional neural networks have shown significant promise in the domain of road scene parsing. Nevertheless, the existing works focus primarily on freespace detection, with little attention given to hazardous road defects that could compromise both driving safety and comfort. In this article, we introduce RoadFormer, a novel Transformer-based data-fusion network developed for road scene parsing. RoadFormer utilizes a duplex encoder architecture to extract heterogeneous features from both RGB images and surface normal information. The encoded features are subsequently fed into a novel heterogeneous feature synergy block for effective feature fusion and recalibration. The pixel decoder then learns multi-scale long-range dependencies from the fused and recalibrated heterogeneous features, which are subsequently processed by a Transformer decoder to produce the final semantic prediction. Additionally, we release SYN-UDTIRI, the first large-scale road scene parsing dataset that contains over 10,407 RGB images, dense depth images, and the corresponding pixel-level annotations for both freespace and road defects of different shapes and sizes. Extensive experimental evaluations conducted on our SYN-UDTIRI dataset, as well as on three public datasets, including KITTI road, CityScapes, and ORFD, demonstrate that RoadFormer outperforms all other state-of-the-art networks for road scene parsing. Specifically, RoadFormer ranks first on the KITTI road benchmark. Our source code, created dataset, and demo video are publicly available at mias.group/RoadFormer.**

***Index Terms*—Convolutional neural network, road scene parsing, freespace detection, semantic segmentation, driving safety and comfort, Transformer.**

I. Introduction

THE advancements in machine intelligence have led to the extensive integration of autonomous driving technologies into various aspects of daily life and across multiple industries [1]. This integration spans a diverse range of products, including autonomous vehicles [2], mobile robots [3], and smart wheelchairs [4]. Recently, researchers in this field have shifted their focus towards enhancing both driving safety and comfort [5]. Road scene parsing, typically including pixel-level freespace and road defect detection, is of paramount importance in achieving these objectives [6].

Existing road scene parsing approaches predominantly fall under two categories: geometry-based and data-driven ones [1]. The algorithms in the former category typically leverage explicit geometric models to represent regions of interest (RoIs), and then proceed to optimize specific energy functions for accurate RoI extraction. For instance, the study presented in [7] employs a B-spline model to fit the road disparity map, which is subsequently projected onto a 2D v-disparity histogram for freespace detection. Additionally, the research presented in [8] introduces a disparity map transformation algorithm designed specifically for the effective detection of road defects. Nonetheless, actual roads are often uneven, rendering such approaches occasionally infeasible [9].

With the proliferation of deep learning techniques, convolutional neural networks (CNNs) have emerged as a potent tool for road scene parsing, often treating it as a binary or ternary semantic segmentation task [1], [10]. These methods have demonstrated significant performance gains over traditional geometry-based approaches. For example, the study detailed in [11] employs an encoder-decoder CNN architecture to realize freespace detection by segmenting RGB images in the bird's eye view. Nonetheless, the results achieved by this approach fall short of satisfactory performance benchmarks. To address this limitation, ensuing research has explored the utilization of data-fusion networks with duplex encoder architectures as a feasible strategy to improve the road scene parsing accuracy. Specifically, [12] extracts heterogeneous features from RGB-Depth data, and subsequently performs feature fusion through a basic element-wise summation operation. The fusion of disparate feature types brings a more comprehensive understanding of the given scenario, resulting in superior performance over earlier single-modal networks. Similar to [12], SNE-RoadSeg series [1], [13] performs RGB-Normal feature fusion through element-wise summation. By employing a duplex ResNet [14] in conjunction with a strong densely-connected decoder, the SNE-RoadSeg series achieves state-of-the-art (SoTA) performance on the KITTI road benchmark [15]. Nevertheless, a current bottleneck exists in the simplistic and indiscriminate fusion of heterogeneous features, often causing conflicting feature representations and erroneous detection results.

Within the domain of computer vision, Transformers have

This research was supported by the National Science and Technology Major Project under Grant 2020AAA0108101, the National Natural Science Foundation of China under Grant 62233013, the Science and Technology Commission of Shanghai Municipal under Grant 22511104500, the Fundamental Research Funds for the Central Universities, and Xiaomi Young Talents Program. (*Corresponding author: Rui Fan*)

Jiahang Li, Yikang Zhang, Guangliang Zhou, Qijun Chen, and Rui Fan are with the College of Electronics & Information Engineering, Shanghai Research Institute for Intelligent Autonomous Systems, the State Key Laboratory of Intelligent Autonomous Systems, and Frontiers Science Center for Intelligent Autonomous Systems, Tongji University, Shanghai 201804, P. R. China (e-mails: {lijiahang617, yikangzhang, tj_zgl, qjchen}@tongji.edu.cn; rui.fan@ieee.org).

Peng Yun is with the Department of Computer Science and Engineering, The Hong Kong University of Science and Technology, Hong Kong. (e-mail: pyun@connect.ust.hk).

Authorized licensed use limited to: TONGJI UNIVERSITY. Downloaded on June 16,2024 at 05:25:01 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Intelligent Vehicles. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIV.2024.3388726

empirically demonstrated their potential to outperform CNNs, particularly when large-scale datasets are available for training [16]–[18]. This advantage can be attributed to the self-attention mechanisms inherent to Transformers, which provide a notably more efficient strategy for global context modeling compared to conventional CNNs [19]. Therefore, employing attention mechanisms to improve the fusion of heterogeneous features extracted by the duplex encoder is an area of research gaining popularity and deserving further attention. OFF-Net [20] is the first attempt to apply the Transformer architecture for road scene parsing. With abundant off-road training data, it outperforms CNN-based algorithms. Unfortunately, OFF-Net utilizes a lightweight CNN-based decoder instead of a Transformer-based one. We believe that the adoption of a Transformer-based decoder has the potential to elevate the upper limit of road scene parsing performance. We also observe its unsatisfactory performance on urban road scenes, especially when the data are limited.

Therefore, in this article, we introduce RoadFormer, a novel duplex Transformer architecture designed for data-fusion semantic road scene parsing. Benefiting from its duplex encoder, RoadFormer can extract abstract and informative heterogeneous features from RGB-Normal data. Additionally, we introduce a novel Heterogeneous Feature Synergy Block (HFSB), which draws upon the self-attention mechanism to improve feature fusion and recalibration. Extensive experiments conducted on public freespace detection datasets demonstrate that RoadFormer achieves improved overall performance than existing SoTA networks. Specifically, RoadFormer ranks first on the KITTI road benchmark upon submission [15]. In contrast to existing studies and their utilized datasets in the domain of road scene parsing [2], [21], [22], which predominantly characterize freespace as undamaged, the aspect of road defects remains notably under-explored. This oversight can be attributed, in part, to the sporadic nature of road defects, making the collection of comprehensive, large-scale datasets containing both RGB and depth images challenging. To this end, we create SYN-UDTIRI, a large-scale, multi-source synthetic dataset, specifically for the understanding of road scenes inclusive of defects. By publishing this dataset, we not only advance the scope of semantic parsing in road scenes but also open up new avenues for data-driven research, thereby demonstrating the capabilities of our proposed RoadFormer.

In summary, the main contributions of this article include:

- RoadFormer, a Transformer-based data-fusion network is developed for semantic road scene parsing, where surface normal information is introduced to bring a more comprehensive understanding of road scenes, our RoadFormer achieves SoTA performance on the KITTI Road benchmark.
- A novel heterogeneous feature synergy block based on self-attention mechanism is designed to dynamically fuse RGB and normal features. Compared to other famous pure CNN-based attention modules, HFSB yields better results when paired with our RoadFormer.
- SYN-UDTIRI, a large-scale synthetic dataset is created for comprehensive road scene parsing. It contains 10,407 pairs of RGB and depth images, as well as pixel-level semantic annotations for freespace and road defects. We believe this dataset and benchmark could bridge the research gap for data-fusion road scene parsing, especially for road defect detection.

The remainder of this article is organized as follows: Sect. II reviews related works. Sect. III details our proposed RoadFormer. Sect. IV presents the experimental results and compares our network with other SoTA single-modal and data-fusion models. Finally, we conclude this article and discuss possible future work in Sect. V.

II. Related Work

A. Single-Modal Networks

Fully convolutional network (FCN) [23] pioneered the utilization of CNNs for single-modal semantic segmentation. Although FCN significantly outperforms traditional methods based on hand-crafted features, its performance is constrained by the absence of multi-scale feature utilization [24]. To address this limitation, DeepLabv3+ [25] employs atrous convolutions with different dilation rates, thereby enhancing the network's capability to encode contextual features across multiple scales. The high-resolution network (HRNet) [26] employs an alternative strategy for multi-scale feature encoding. Instead of solely performing detrimental resizing of feature maps, HRNet integrates low-resolution branches in parallel with the high-resolution main branch to achieve multi-scale feature extraction.

Transformers have been applied to semantic segmentation tasks due to their superior global aggregation capabilities over CNNs [27]. Segmentation Transformer (SETR) [28] is the first Transformer-based general-purpose semantic segmentation network. Building on the success of Vision Transformer (ViT) [29], it tokenizes images into patches and feeds them into Transformer blocks. These encoded features are then gradually restored through upsampling convolution to achieve pixel-level classification. SegFormer [18] introduces a multi-scale Transformer encoder for semantic segmentation, which stacks Transformer blocks and inserts convolutional layers (for feature map downsampling) between them. Compared to SETR, SegFormer improves segmentation performance when dealing with objects of varying sizes. Additionally, MaskFormer [30] introduces a new paradigm for semantic segmentation. Rather than adhering to standard per-pixel classification methods, this architecture addresses the semantic segmentation task by decoding query features into a set of masks. In detail, MaskFormer employs a multi-scale Transformer decoder that outputs the mask for each class using refined queries in parallel, outperforming previous per-pixel classification approaches. This query-based prediction fashion is also adopted in our proposed method and has demonstrated improved performance over SoTA CNNs in terms of segmentation accuracy.

B. Data-Fusion Networks

Data-fusion networks effectively leverage heterogeneous features extracted from RGB images and spatial geometric information to improve segmentation performance. FuseNet

Authorized licensed use limited to: TONGJI UNIVERSITY. Downloaded on June 16,2024 at 05:25:01 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Intelligent Vehicles. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIV.2024.3388726

[12] was the first attempt to incorporate depth information into semantic segmentation, using separate CNN encoders for RGB and depth images and fusing their features through element-wise summation. MFNet [31] strikes a balance between speed and accuracy in driving scene parsing through RGB-Thermal fusion. Similarly, RTFNet [32] also uses RGB-Thermal data as inputs and designs a robust decoder that utilizes shortcuts to produce clear boundaries while retaining detailed features. Inspired by [12], SNE-RoadSeg [1] and SNE-RoadSeg+ [13] incorporate surface normal information into freespace detection. This series employs densely connected skip connections to enhance feature extraction in their decoder, thereby achieving SoTA performance compared to other approaches. Drawing on the success of single-modal Transformer models, OFF-Net [20] was the first attempt to apply Transformer architecture for data-fusion freespace detection. It utilizes a SegFormer [18] encoder to generate RGB and surface normal features, outperforming SoTA CNNs in off-road freespace detection. Expanding upon these foundational prior arts, our RoadFormer also adopts the data-fusion paradigm but differentiates itself by employing a novel Transformer architecture for semantic road scene parsing. Additionally, RoadFormer utilizes a novel feature synergy block, leading to superior performance over all other data-fusion networks on four road scene parsing datasets.

III. Methodology

This section details RoadFormer, a robust and powerful data-fusion Transformer architecture for road scene parsing. As depicted in Fig. 1, RoadFormer consists of:

(1) a duplex encoder to learn heterogeneous features from RGB-Normal data;
(2) a feature synergy block to fuse and recalibrate the encoded heterogeneous features;
(3) a pixel decoder to learn long-range dependencies from recalibrated features;
(4) a Transformer decoder to achieve final semantic prediction by refining query features using outputs of the pixel decoder.

A. Duplex Encoder

In line with our previous works [1], [4], [13], we employ a duplex encoder structure to extract multi-scale heterogeneous features. One encoder focuses on learning color and texture features $\mathcal{F}^R = \{\boldsymbol{F}_1^R, \ldots, \boldsymbol{F}_k^R\}$ from RGB images $\boldsymbol{I}^R \in \mathbb{R}^{H\times W\times 3}$, while the other encoder specializes in acquiring representations $\mathcal{F}^N = \{\boldsymbol{F}_1^N, \ldots, \boldsymbol{F}_k^N\}$ of planar characteristics from the surface normal information $\boldsymbol{I}^N \in \mathbb{R}^{H\times W\times 3}$, where H and W represent the height and width of the input image, respectively, $\boldsymbol{F}_i^{R,N} \in \mathbb{R}^{\frac{H}{S_i}\times\frac{W}{S_i}\times C_i}$ represents the i-th feature maps, and C_i and $S_i = 2^{i+1}$ ($i \in [1,k] \cap \mathbb{Z}$) denote the corresponding channel and stride numbers, respectively (usually $k = 4$). Our duplex encoder is compatible with both Transformer-based and CNN-based backbones. For this study, we utilize Swin Transformer [17] and ConvNeXt [33] as our backbones, respectively. The performance comparison of these backbones is presented in Sect. IV-C.

B. Heterogeneous Feature Synergy Block

1) Heterogeneous Feature Fusion Module (HFFM): Conventional duplex networks typically fuse heterogeneous features using basic element-wise addition or feature concatenation operations [4]. Nevertheless, we contend that such a simplistic feature fusion strategy might not fully exploit the inherent potential of heterogeneous features. The recent surge in the popularity of Transformer architectures for multi-modal visual-linguistic tasks [34], [35] can be attributed to their powerful attention mechanisms that enable effective joint representations across modalities [36]. Drawing inspiration from these prior arts, our objective is to utilize attention mechanisms to enhance the fusion of heterogeneous features extracted by the duplex encoder mentioned above. To this end, we introduce HFFM (see Fig. 2 (a)), which builds upon the self-attention mechanisms of Transformers to achieve effective fusion of $\mathcal{F}^R$ and $\mathcal{F}^N$ through token interactions. Our HFFM can be formulated as follows:

$$\boldsymbol{F}_i^H = \mathrm{Reshape}\Big(\mathrm{Norm}\big(\mathrm{Softmax}(\boldsymbol{Q}_i^C \boldsymbol{K}_i^{C\top})\kappa_i \boldsymbol{V}_i^C + \boldsymbol{F}_i^C\big)\Big), \tag{1}$$

where $\boldsymbol{F}_i^R$ and $\boldsymbol{F}_i^N$ are concatenated and reshaped to form $\boldsymbol{F}_i^C \in \mathbb{R}^{2C_i\times\frac{H}{S_i}\frac{W}{S_i}}$, which is then identically mapped to query $\boldsymbol{Q}_i^C$, key $\boldsymbol{K}_i^C$, and value $\boldsymbol{V}_i^C$ embeddings, and $\boldsymbol{F}_i^H \in \mathbb{R}^{\frac{H}{S_i}\times\frac{W}{S_i}\times 2C_i}$ denotes the output of the HFFM. Additionally, following [37], we introduce a learnable coefficient κ_i to adaptively adjust the attention significance, enabling a more flexible fusion of heterogeneous features.

2) Fused Feature Recalibration Module (FFRM): In conventional single-encoder architecture designs, multi-channel features do not all contribute positively to semantic predictions. In fact, some noisy or irrelevant feature maps may even degrade the model's performance. In this regard, the squeeze-and-excitation block (SEB) [38] was designed to model the inter-dependencies between the channels of convolutional features via a channel attention mechanism. This allows the network to focus on more informative features while downplaying the less important ones. A similar issue also occurs in our work: the heterogeneous features extracted by our duplex encoder may focus on different components in the scene, and the fusion of these features might potentially undermine the saliency of the original key features or even produce irrelevant features [39]. To this end, we develop FFRM (see Fig. 2 (b)) based on SEB to recalibrate the fused heterogeneous features. In particular, we add a residual connection atop the SEB to enhance its training [14] and introduce an additional point-wise convolution to realize the flexible computation of correlations between the recalibrated features [40]. Our FFRM can be formulated as follows:

$$\boldsymbol{F}_i^F = \underset{1\times1}{\mathrm{Conv}}\left(\boldsymbol{F}_i^H + \Big(\boldsymbol{O}\ \mathrm{Sigmoid}\big(\underset{1\times1}{\mathrm{Conv}}(\boldsymbol{z}_i)\big)\Big) \odot \boldsymbol{F}_i^H\right), \tag{2}$$

where $\boldsymbol{O} \in \mathbb{R}^{\frac{H}{S_i}\times\frac{W}{S_i}\times 1\times 1}$ represents a matrix of ones, $\odot$ denotes the Hadamard product operation, $\boldsymbol{z}_i = [z_{i,1}, \ldots, z_{i,2C_i}] \in \mathbb{R}^{1\times1\times2C_i}$ stores the average pooling

Authorized licensed use limited to: TONGJI UNIVERSITY. Downloaded on June 16,2024 at 05:25:01 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Intelligent Vehicles. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIV.2024.3388726

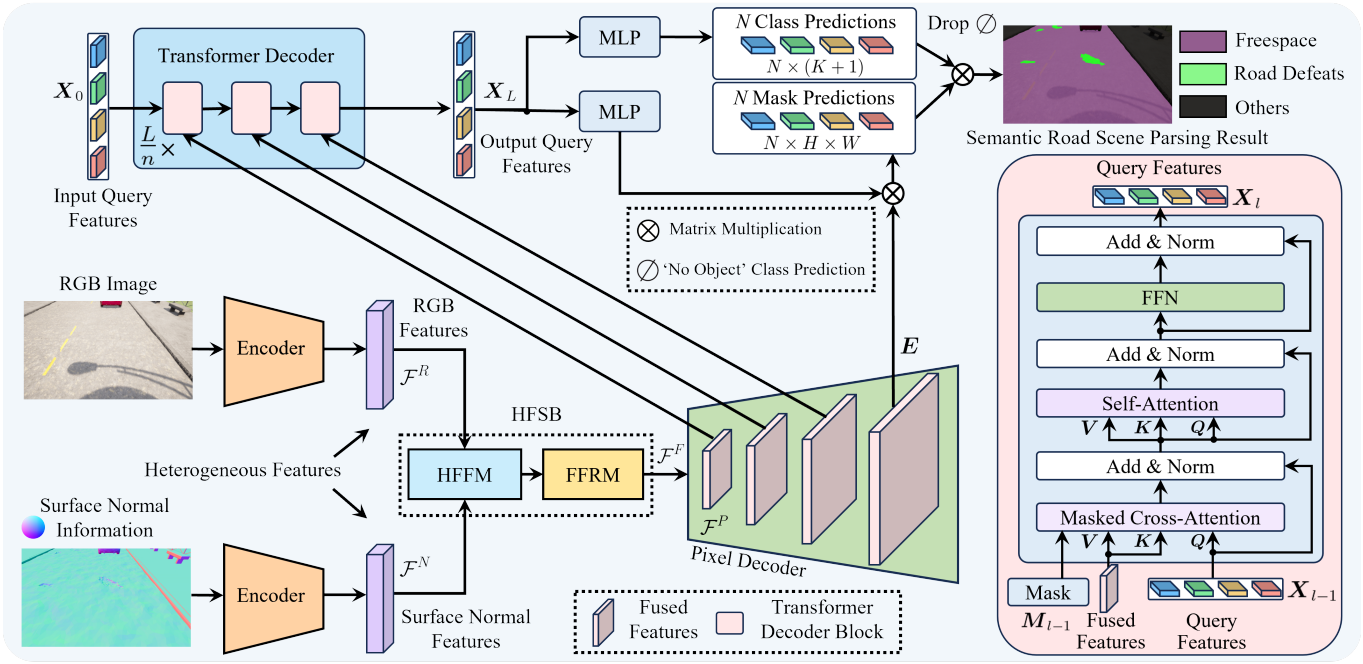


Fig. 1. An overview of our proposed RoadFormer architecture.

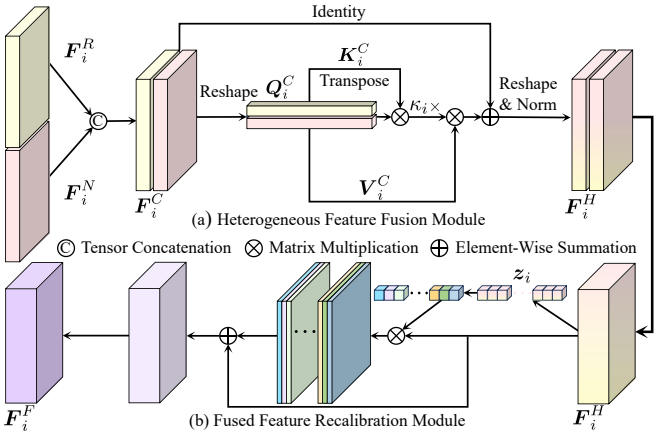


Fig. 2. Heterogeneous feature synergy block.

results of each feature map in $\boldsymbol{F}_i^H$, and

$$z_{i,j} = \frac{S_i^2}{HW} \sum_{h=1}^{\frac{H}{S_i}} \sum_{w=1}^{\frac{W}{S_i}} \boldsymbol{F}_i^H(h, w, j). \tag{3}$$

The performance comparison between our proposed HFFM, FFRM, and SE block is discussed in Sect. II.

C. Pixel Decoder

Following [41], we incorporate a pixel decoder to improve multi-scale feature modeling for $\mathcal{F}^F = \{\boldsymbol{F}_1^F, \ldots, \boldsymbol{F}_k^F\}$ to generate $\mathcal{F}^P = \{\boldsymbol{F}_1^P, \ldots, \boldsymbol{F}_n^P\}$ ($n < k$, usually $n = 3$), where $\boldsymbol{F}_i^P \in \mathbb{R}^{\frac{H}{S_i} \times \frac{W}{S_i} \times C}$. It also serves the function of upsampling low-resolution features in $\mathcal{F}^F$ to generate a high-resolution per-pixel embedding $\boldsymbol{E}$, which is then used by the Transformer decoder to guide the mask prediction. Given the recent success of multi-scale deformable attention Transformer [41]–[43], we adopt this architecture within our pixel decoder to generate $\mathcal{F}^P$. In contrast to the single-modal approach presented in [41], our pixel decoder processes the fused features $\mathcal{F}^F$ generated from both RGB images and surface normal information. Consequently, the feature map channel at each scale is doubled compared to [41]. To reduce the increased computational complexity, we employ 1×1 convolutional layers to reduce the feature map channels before performing multi-scale deformable attention operations.

D. Transformer Decoder

We utilize a Transformer-based decoder to recursively update the input query features $\boldsymbol{X}_0 \in \mathbb{R}^{N \times C}$ based on the multi-scale feature maps $\boldsymbol{F}_1^P$ to $\boldsymbol{F}_n^P$ extracted by the pixel decoder. The per-pixel embedding $\boldsymbol{E}$ is used to guide mask predictions. Specifically, N learnable feature vectors with C channels are initialized as the input query features $\boldsymbol{X}_0$, which are fed into the subsequent Transformer decoder layers. One Transformer decoder layer consists of a sequence of operations: (1) the query $\boldsymbol{Q}_l^D = f_Q(\boldsymbol{X}_{l-1}) \in \mathbb{R}^{N \times C}$ is obtained through a linear transformation operation $f_Q(\cdot)$, where l is the layer index; (2) the key $\boldsymbol{K}_l^D = f_K(\boldsymbol{F}_i^P) \in \mathbb{R}^{\frac{H}{S_i}\frac{W}{S_i} \times C}$ and value $\boldsymbol{V}_l^D = f_V(\boldsymbol{F}_i^P) \in \mathbb{R}^{\frac{H}{S_i}\frac{W}{S_i} \times C}$ are obtained through two linear transformation operations $f_K(\cdot)$ and $f_V(\cdot)$, respectively; (3) $\boldsymbol{Q}_l^D$, $\boldsymbol{K}_l^D$, and $\boldsymbol{V}_l^D$ are subsequently processed by a masked cross-attention mechanism, expressed as follows:

$$\boldsymbol{X}_l^C = \mathrm{Softmax}(\boldsymbol{M}_{l-1} + \boldsymbol{Q}_l^D \boldsymbol{K}_l^{D\top}) \boldsymbol{V}_l^D + \boldsymbol{X}_{l-1}, \tag{4}$$

where $\boldsymbol{M}_{l-1} \in \mathbb{R}^{N \times \frac{H}{S_i}\frac{W}{S_i}}$ (containing only the values of 0 or $-\infty$) denotes the output of the resized mask prediction obtained from the $(l-1)$-th Transformer decoder layer (readers

Authorized licensed use limited to: TONGJI UNIVERSITY. Downloaded on June 16,2024 at 05:25:01 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Intelligent Vehicles. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIV.2024.3388726

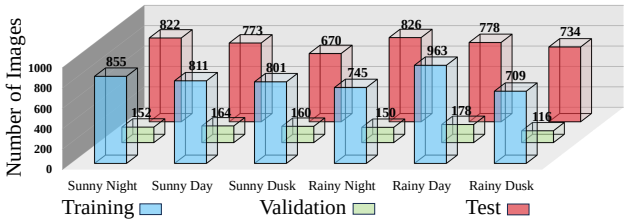


Fig. 3. Data distribution in the SYN-UDTIRI dataset.

can refer to [41] for more details); (4) $\boldsymbol{X}_l^C$, the masked cross-attention results are subsequently processed by a self-attention mechanism, and ultimately passed through a feedforward network (FFN) to generate the query features $\boldsymbol{X}_l$. Multi-scale feature maps $\boldsymbol{F}_1^P$ to $\boldsymbol{F}_n^P$ are fed to their corresponding decoder layer (n successive layers), and the entire process is iterated $\frac{L}{n}$ times to update the query features. Following [44], the output query features $\boldsymbol{X}_L$ produced by the Transformer decoder is mapped into a space of dimension $(K+1)$ for class predictions by a multi-layer perceptron (MLP), where K represents the total number of classes to be predicted, plus an additional class representing "no object". The mask prediction is then obtained by performing a dot product operation between the mask embedding $\mathrm{MLP}(\boldsymbol{X}_L)$ (also generated by MLP) and the per-pixel embedding $\boldsymbol{E}$. Finally, the semantic road scene parsing result is obtained by performing a simple matrix multiplication operation (followed by an argmax function) between the mask and class predictions. Notably, this query-based decoder pipeline is identical to the approach proposed in [30], where semantic predictions are obtained from $\boldsymbol{X}_L$. Each query feature in $\boldsymbol{X}_L$ generates one specific mask prediction and the corresponding class predictions for $K+1$ classes, collectively forming the final semantic segmentation result.

E. Loss Function

We follow [41] to train our proposed RoadFormer by minimizing the following loss function:

$$\mathcal{L} = \lambda_{\text{mask}}(\lambda_{\text{ce}}\mathcal{L}_{\text{ce}} + \lambda_{\text{dice}}\mathcal{L}_{\text{dice}}) + \lambda_{\text{cls}}\mathcal{L}_{\text{cls}}, \tag{5}$$

where $\mathcal{L}_{\text{ce}}$ and $\mathcal{L}_{\text{dice}}$ are the binary cross-entropy loss and the dice loss, respectively. $\mathcal{L}_{\text{cls}}$ is the classification loss, and the weighting factors λ_{mask}, λ_{ce}, λ_{dice}, and λ_{cls} serve to balance the respective contributions of the different loss components to the overall loss. These are set in accordance with [41].

IV. Experiments

A. Datasets

1) ***SYN-UDTIRI****:* Owing to the lack of well-annotated, large-scale datasets, created specifically for road scene parsing (freespace and road defect detection), we create a synthetic dataset, referred to as SYN-UDTIRI, using the CARLA simulator [45]. The principal contribution of this dataset lies in the integration of digital twins of road potholes acquired from the real world using our previously published 3D geometry reconstruction algorithm [46], [47]. Moreover, to better simulate the roughness of actual roads, we introduce random Perlin noise [48] to the road data. We generate six driving scenarios, including sunny day, dusk, and night, as well as rainy day, dusk, and night, with respect to different illumination and weather conditions. Additionally, we deploy a simulated stereo rig (baseline: 0.5 m) onto a moving vehicle to acquire over 10K pairs of stereo road images (resolution: $720 \times 1,280$ pixels), along with their corresponding depth images, surface normal information, and semantic annotations, including three categories: freespace, road defect, and other objects. More details on the SYN-UDTIRI dataset are given in Fig. 3.

2) ***KITTI Road****:* The KITTI road [15] dataset contains 289 pairs of stereo images and their corresponding LiDAR point clouds for model training and validation. It also provides a comparable amount of testing data without semantic annotations. We follow a similar data pre-processing strategy as detailed in [1]. We fine-tune our proposed RoadFormer for the test set result submission to the KITTI road benchmark.

3) ***CityScapes****:* The CityScapes [22] dataset is a widely utilized urban scene dataset, containing 2,975 stereo training images and 500 validation images, with well-annotated semantic annotations. Due to the limited sample size of the KITTI road dataset, we conduct additional experiments on the CityScapes dataset to further demonstrate the effectiveness of our proposed RoadFormer on large-scale datasets. All experimental results are obtained using the validation set since ground-truth annotations are not provided on the test set. Quantitative and qualitative evaluation results on the test set are typically acquired by submitting results to the online CityScapes benchmark suite. Furthermore, given our specific focus on road scene parsing, we have to reorganize the dataset to train and evaluate models for only two classes: road and others. It is noteworthy that the corresponding surface normal information is derived from depth images obtained using RAFT-Stereo [49] trained on the KITTI [50] dataset.

4) ***ORFD****:* The ORFD [20] dataset is designed specifically for off-road freespace detection. It contains 12,198 RGB images and their corresponding LiDAR point clouds, collected across various scenes, under different weather and illumination conditions. We follow the data splitting and pre-processing strategies (except for surface normal estimation) detailed in [20] for our experiments.

B. Experimental Setup and Evaluation Metrics

In our experiments, we compare our proposed RoadFormer with four single-modal networks and five data-fusion networks. The single-modal networks are trained using RGB images, while the data-fusion networks are trained using both RGB images and surface normal information (estimated using D2NT [51] owing to its superior accuracy compared to other methods). All the networks undergo training for the same number of epochs. For RoadFormer training, we utilize the AdamW optimizer [52] with a polynomial learning rate decay strategy [53]. The learning rate begins at 10^{-4} with a weight decay of 5×10^{-2}. We apply learning rate multipliers of 10^{-1} to both ConvNeXt [33] and Swin [17] backbones.

We employ five common metrics to quantify the network performance: accuracy (Acc), precision (Pre), recall (Rec),

Authorized licensed use limited to: TONGJI UNIVERSITY. Downloaded on June 16,2024 at 05:25:01 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Intelligent Vehicles. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIV.2024.3388726

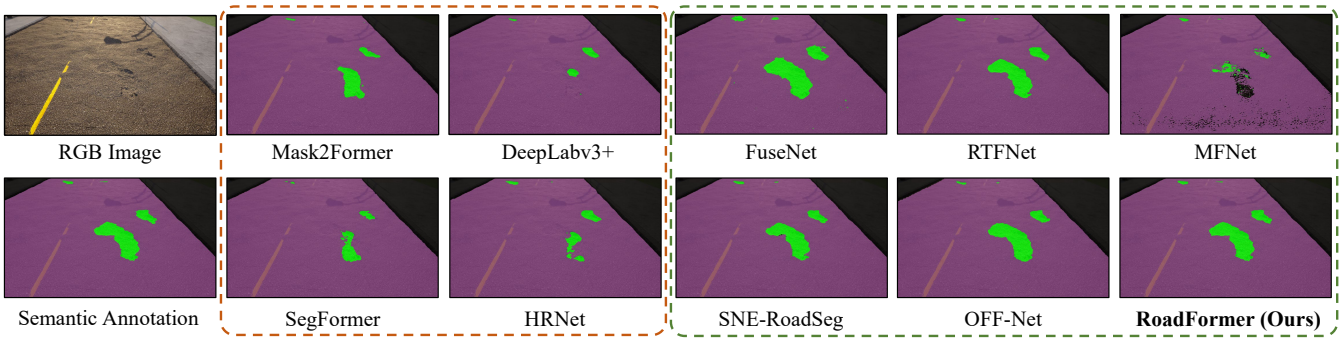


Fig. 4. Qualitative comparison between RoadFormer and other SoTA networks on the SYN-UDTIRI dataset. The freespace, road defect, and background areas are shown in purple, green, and black, respectively.

intersection over union (IoU), and F-score (Fsc). We refer readers to our previous work [1] for more details on these metrics. Additionally, the evaluation metrics used for the KITTI road benchmark can be found on its official webpage: https://www.cvlibs.net/datasets/kitti/eval_road.php.

TABLE I
ABLATION STUDY ON BACKBONE SELECTION.

Dataset	Backbone	IoU (%)	Fsc (%)	Pre (%)	Rec (%)
SYN-UDTIRI	ConvNeXt	**93.38**	**96.58**	96.58	**96.74**
	Swin	93.18	96.47	**96.59**	96.35
CityScapes	ConvNeXt	**95.80**	**97.86**	**97.74**	**97.97**
	Swin	94.69	97.27	97.25	97.29

TABLE II
ABLATION STUDY DEMONSTRATING THE EFFECTIVENESS OF OUR PROPOSED HFFM AND FFRM.

SEB	HFFM	FFRM	IoU (%)	Acc (%)	Fsc (%)	Pre (%)	Rec (%)	FPS
×	×	×	95.11	96.87	97.50	98.12	96.87	21.80
✓	×	×	95.45	97.40	97.67	**97.95**	97.40	21.60
×	×	✓	95.49	97.59	97.69	97.79	97.59	21.60
×	✓	×	95.34	97.67	97.61	97.55	97.67	20.50
×	✓	✓	**95.80**	**97.97**	**97.86**	97.74	**97.97**	20.10

C. Ablation Study

Although recent studies on scene parsing, such as [41] and [43], often combine the widely used Swin Transformer encoder with a Transformer decoder, we hypothesize that the Swin Transformer might not be the optimal choice for all parsing tasks. Therefore, we perform an ablation study on SYN-UDTIRI and CityScapes to compare the performance of ConvNeXt and Swin Transformer for encoder backbone selection. As shown in Table I, ConvNeXt demonstrates superior performance over Swin, achieving improvements of 0.20%-1.11% in terms of IoU and 0.11%-0.59% in terms of F-score. These results suggest that our proposed RoadFormer is compatible with both CNN-based and Transformer-based backbones, and ConvNeXt is generally a preferable option for road scene parsing. Therefore, in the following experiments, we will utilize ConvNeXt as our backbone network.

Given that the design of FFRM is inspired by SEB, we first compare the performance of our RoadFormer when incorporated with either SEB or FFRM. Furthermore, we also evaluate the individual effectiveness of HFFM and FFRM, as well as their compatibility. The results shown in Table II indicate that (1) FFRM outperforms SEB, achieving an improvement of 0.04% in terms of IoU, (2) HFFM yields a 0.23% higher IoU compared to the baseline setup, and (3) the combined utilization of HFFM and FFRM modules results in better performance than using these modules independently.

Despite our HFSB introducing a slight overhead on inference speed due to its multi-scale recalibration and fusion of heterogeneous features, RoadFormer achieves an inference speed of ~ 20 FPS when processing images at a resolution of 352×640 pixels on an NVIDIA RTX 3090 GPU. This performance effectively meets real-time processing requirements.

D. Experimental Results

TABLE III
COMPARISON OF SOTA ALGORITHMS PUBLISHED ON THE KITTI ROAD BENCHMARK.

Method	MaxF (%)	AP (%)	Pre (%)	Rec (%)	Rank
NIM-RTFNet [54]	96.02	94.01	96.43	95.62	13
HID-LS [55]	93.11	87.33	92.52	93.71	33
LC-CRF [56]	95.68	88.34	93.62	97.83	15
SNE-RoadSeg [1]	96.75	94.07	96.90	96.61	8
SNE-RoadSeg+ [13]	97.50	93.98	**97.41**	97.58	2
PLB-RD [57]	97.42	**94.09**	97.30	97.54	3
LRDNet+ [58]	96.95	92.22	96.88	97.02	4
DFM-RTFNet [4]	96.78	94.05	96.62	96.93	7
RoadFormer	**97.50**	93.85	97.16	**97.84**	**1**

The quantitative results on the SYN-UDTIRI, CityScapes, ORFD, and KITTI Road datasets are presented in Tables III-VI. Additionally, we also present readers with the qualitative results on these four datasets in Figs. 4-7. These results suggest that our proposed RoadFormer outperforms all other SoTA networks across all four datasets, demonstrating its exceptional

Authorized licensed use limited to: TONGJI UNIVERSITY. Downloaded on June 16,2024 at 05:25:01 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Intelligent Vehicles. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIV.2024.3388726

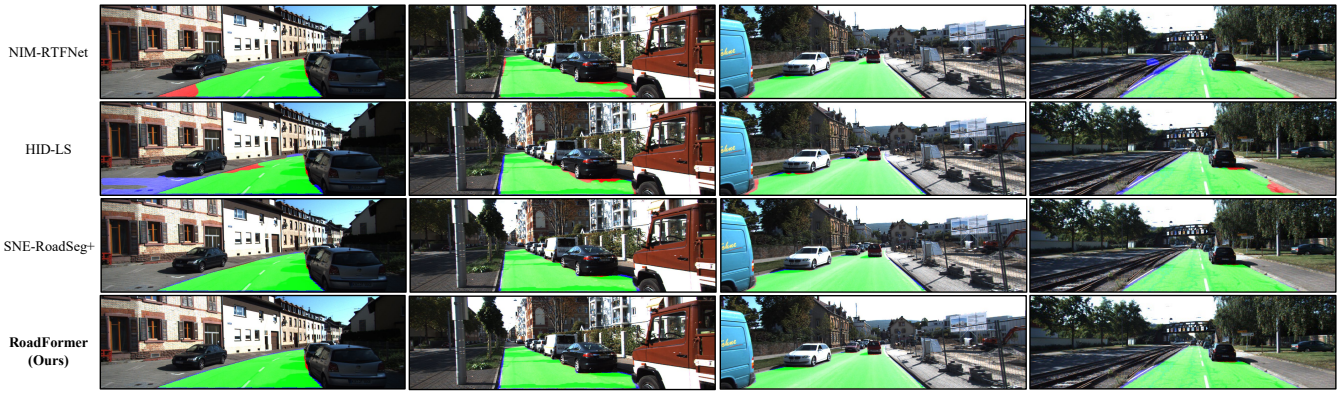


Fig. 5. Qualitative comparison between RoadFormer and other SoTA networks on the KITTI road test set. The results are obtained from the official KITTI online benchmark suite. True-positive, false-positive, and false-negative classifications are shown in green, blue, and red, respectively.

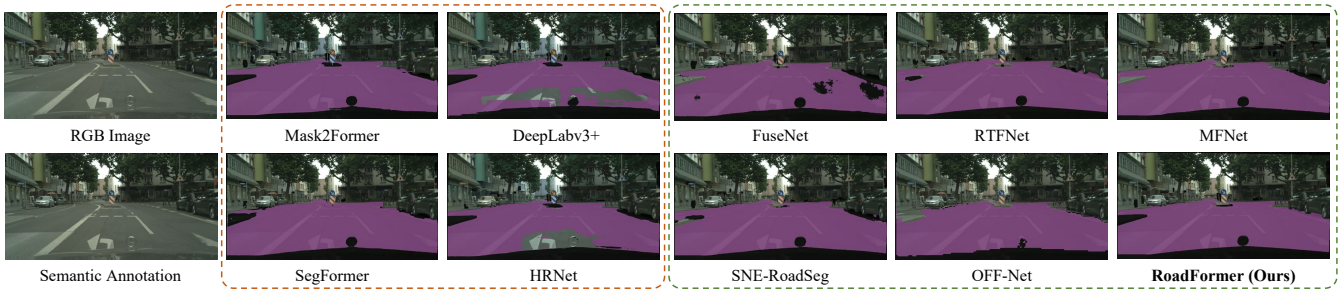


Fig. 6. Qualitative comparison between RoadFormer and other SoTA networks on the CityScapes dataset. The freespace and ignored areas are shown in purple and black, respectively.

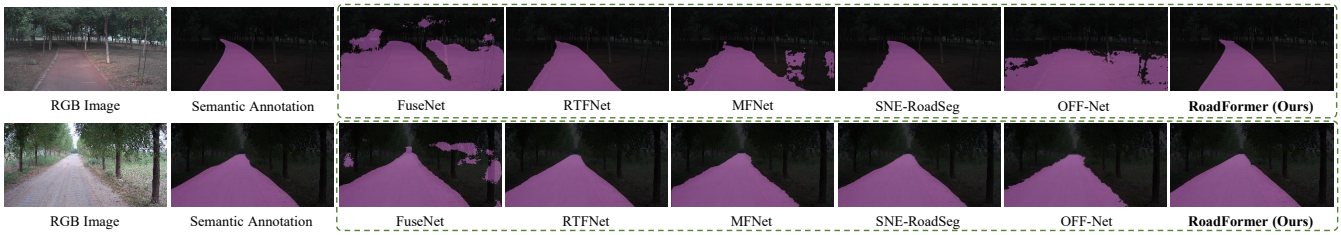


Fig. 7. Qualitative comparison between RoadFormer and other SoTA data-fusion networks on the ORFD dataset. The freespace and remaining areas are shown in purple and black, respectively.

performance and robustness in effectively parsing various types of road scenes, including synthetic roads with defects, urban roads, and rural roads.

As shown in Table IV, it is evident that data-fusion networks demonstrate considerably improved robustness compared to single-modal networks, particularly in road defect detection. We exclude the freespace detection results for comparison due to the closely matched performance across all networks in this task (the single-modal and data-fusion networks achieve IoUs of over 98.5% and 99.5%, respectively). As expected, data-fusion networks achieve better generalizability compared to single-modal networks. This improvement is attributed to geometric features extracted from surface normal information.

Additionally, the results on the CityScapes dataset somewhat exceed our expectations. The performance of the single-modal networks, with the IoU fluctuating within a range of 0.2%, is quite comparable to that of the data-fusion networks. To our surprise, all single-modal networks outperform data-fusion networks, except for RTFNet and our RoadFormer. We speculate that these unexpected results could be attributed to inaccuracies in the disparity maps used for surface normal estimation, as they are directly obtained from pre-trained stereo matching networks. The previous data-fusion networks employ basic element-wise addition or feature concatenation operations for feature fusion, leading to performance degradation when surface normal information is inaccurate. In contrast, benefiting from the adaptive fusion and recalibration of heterogeneous features via our designed HFSB, our RoadFormer achieves the highest scores across all metrics, including the mean IoU (mIoU) computed across 20 categories (including the "ignore" category) in the full-pixel semantic segmentation task.

To ensure a fair comparison with SoTA networks on the ORFD dataset, we present both the results reported in the original paper [20] for three networks and those obtained through our re-implementation for six networks, as shown in Table

Authorized licensed use limited to: TONGJI UNIVERSITY. Downloaded on June 16,2024 at 05:25:01 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Intelligent Vehicles. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIV.2024.3388726

TABLE IV
QUANTITATIVE RESULTS ON THE SYN-UDTIRI DATASET.

	Subset	Method	IoU (%)	Fsc (%)	Pre (%)	Rec (%)
RGB	Validation	Mask2Former	64.29	78.27	83.0	74.05
		SegFormer	52.46	68.82	70.13	67.55
		DeepLabv3+	52.94	69.23	75.23	64.12
		HRNet	52.92	69.21	79.46	61.30
	Test	Mask2Former	46.91	63.87	73.59	56.41
		SegFormer	36.34	53.31	57.23	49.89
		DeepLabv3+	34.76	51.58	62.54	43.90
		HRNet	35.47	52.37	69.09	42.16
RGB-Normal	Validation	FuseNet	67.30	80.40	68.30	**97.80**
		SNE-RoadSeg	92.00	95.80	96.30	95.40
		RTFNet	90.30	94.90	94.10	95.70
		OFF-Net	83.90	91.30	91.70	90.80
		MFNet	89.50	94.50	95.70	93.30
		RoadFormer	**93.35**	**96.56**	**96.53**	96.59
	Test	FuseNet	70.70	82.90	72.10	**97.50**
		SNE-RoadSeg	92.10	95.90	**96.70**	95.10
		RTFNet	90.50	95.00	95.50	94.50
		OFF-Net	83.80	91.20	91.90	90.50
		MFNet	87.70	93.50	96.20	90.90
		RoadFormer	**93.51**	**96.65**	96.61	96.69

TABLE V
QUANTITATIVE RESULTS ON THE CITYSCAPES DATASET.

	Method	IoU (%)	Fsc (%)	Pre (%)	Rec (%)	mIoU (%)
RGB	Mask2Former	93.84	96.82	97.14	96.51	74.80
	SegFormer	93.98	96.90	96.02	97.79	64.51
	DeepLabv3+	93.82	96.81	96.99	96.63	68.66
	HRNet	94.06	96.94	96.29	97.59	70.10
RGB-Normal	FuseNet	91.60	95.60	96.00	95.30	52.70
	SNE-RoadSeg	93.80	96.80	96.10	97.50	53.40
	RTFNet	94.10	96.90	96.30	97.60	49.60
	OFF-Net	89.60	94.50	93.40	95.70	39.20
	MFNet	92.10	95.90	94.10	97.70	49.30
	RoadFormer	**95.80**	**97.86**	**97.74**	**97.97**	**76.20**

VI. We observe that the performances of data-fusion networks differ significantly on the ORFD dataset. This is likely due to the challenge of labeling accurate off-road semantic ground truth. The inaccurate annotations may introduce ambiguities during the model training process. Finally, we submit the test set results produced by RoadFormer to the KITTI road online benchmark for performance comparison. As shown in Table III, RoadFormer demonstrates superior performance compared to all previously published methods.

V. CONCLUSION AND FUTURE WORK

This article presented RoadFormer, a powerful data-fusion Transformer architecture designed for road scene parsing. It contains a duplex encoder, a novel feature synergy block, and Transformer-based decoders. Compared to previous works,

TABLE VI
QUANTITATIVE RESULTS ON THE TEST SET OF ORFD DATASET. WE PRESENT BOTH THE RESULTS REPORTED IN THE ORIGINAL PAPER [20] AND THOSE OBTAINED THROUGH OUR RE-IMPLEMENTATION. † DENOTES THE MODEL TRAINED USING RGB-DEPTH DATA IN THE ORIGINAL IMPLEMENTATION.

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

RoadFormer demonstrates the effective fusion of heterogeneous features and improved accuracy. It outperforms all existing semantic segmentation networks on our newly created SYN-UDTIRI dataset and three public datasets, while ranking first on the KITTI road benchmark upon submission. As critical components of RoadFormer, feature fusion using self-attention has proven superior to pure CNNs for road scene parsing, and we aim to investigate its potential for more common scene parsing tasks in the future. On the other hand, although achieving high accuracy, the real-time performance of data-fusion networks still needs improvement, which we will leave to future work.

REFERENCES

[1] R. Fan *et al.*, "SNE-RoadSeg: Incorporating Surface Normal Information into Semantic Segmentation for Accurate Freespace Detection," in *European Conference on Computer Vision (ECCV)*. Springer, 2020, pp. 340–356.

[2] A. Geiger *et al.*, "Vision meets robotics: The KITTI dataset," *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1231–1237, 2013.

[3] J. Li *et al.*, "Towards Broad Learning Networks on Unmanned Mobile Robot for Semantic Segmentation," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 9228–9234.

[4] H. Wang *et al.*, "Dynamic Fusion Module Evolves Drivable Area and Road Anomaly Detection: A Benchmark and Algorithms," *IEEE Transactions on Cybernetics*, vol. 52, no. 10, pp. 10 750–10 760, 2021.

[5] Y. Du *et al.*, "Velocity Control Strategies to Improve Automated Vehicle Driving Comfort," *IEEE Intelligent Transportation Systems Magazine*, vol. 10, no. 1, pp. 8–18, 2018.

[6] B. Barabino *et al.*, "Standing Passenger Comfort: A New Scale for Evaluating the Real-Time Driving Style of Bus Transit Services," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 12, pp. 4665–4678, 2019.

[7] A. Wedel *et al.*, "B-Spline Modeling of Road Surfaces With an Application to Free-Space Estimation," *IEEE Transactions on Intelligent Transportation Systems*, vol. 10, no. 4, pp. 572–583, 2009.

[8] R. Fan *et al.*, "Pothole Detection Based on Disparity Transformation and Road Surface Modeling," *IEEE Transactions on Image Processing*, vol. 29, pp. 897–908, 2019.

[9] N. Ma *et al.*, "Computer vision for road imaging and pothole detection: a state-of-the-art review of systems and algorithms," *Transportation safety and Environment*, vol. 4, no. 4, p. tdac026, 2022.

[10] R. Fan *et al.*, "Graph Attention Layer Evolves Semantic Segmentation for Road Pothole Detection: A Benchmark and Algorithms," *IEEE Transactions on Image Processing*, vol. 30, pp. 8144–8154, 2021.

Authorized licensed use limited to: TONGJI UNIVERSITY. Downloaded on June 16,2024 at 05:25:01 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Intelligent Vehicles. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIV.2024.3388726

[11] C. Lu *et al.*, "Monocular Semantic Occupancy Grid Mapping With Convolutional Variational Encoder–Decoder Networks," *IEEE Robotics and Automation Letters*, vol. 4, no. 2, pp. 445–452, 2019.

[12] C. Hazirbas *et al.*, "FuseNet: Incorporating Depth into Semantic Segmentation via Fusion-Based CNN Architecture," in *13th Asian Conference on Computer Vision (ACCV)*. Springer, 2017, pp. 213–228.

[13] H. Wang *et al.*, "SNE-RoadSeg+: Rethinking Depth-Normal Translation and Deep Supervision for Freespace Detection," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021, pp. 1140–1145.

[14] 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.

[15] J. Fritsch *et al.*, "A New Performance Measure and Evaluation Benchmark for Road Detection Algorithms," in *16th International IEEE Conference on Intelligent Transportation Systems (ITSC) 2013*. IEEE, 2013, pp. 1693–1700.

[16] K. Han *et al.*, "A Survey on Vision Transformer," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 1, pp. 87–110, 2022.

[17] Z. Liu *et al.*, "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 10 012–10 022.

[18] E. Xie *et al.*, "SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 34, pp. 12 077–12 090, 2021.

[19] K. Li *et al.*, "UniFormer: Unifying Convolution and Self-Attention for Visual Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[20] C. Min *et al.*, "ORFD: A Dataset and Benchmark for Off-Road Freespace Detection," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 2532–2538.

[21] P. Sun *et al.*, "Scalability in Perception for Autonomous Driving: Waymo Open Dataset," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 2446–2454.

[22] M. Cordts *et al.*, "The Cityscapes Dataset for Semantic Urban Scene Understanding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 3213–3223.

[23] J. Long *et al.*, "Fully Convolutional Networks for Semantic Segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 3431–3440.

[24] H. Ding *et al.*, "Context Contrasted Feature and Gated Multi-Scale Aggregation for Scene Segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 2393–2402.

[25] L. 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.

[26] J. Wang *et al.*, "Deep High-Resolution Representation Learning for Visual Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 10, pp. 3349–3364, 2020.

[27] R. Strudel *et al.*, "Segmenter: Transformer for Semantic Segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 7262–7272.

[28] 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.

[29] A. Dosovitskiy *et al.*, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," *International Conference on Learning Representations (ICLR)*, 2020.

[30] B. Cheng *et al.*, "Per-Pixel Classification is Not All You Need for Semantic Segmentation," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 34, pp. 17 864–17 875, 2021.

[31] Q. Ha *et al.*, "MFNet: Towards Real-Time Semantic Segmentation for Autonomous Vehicles with Multi-Spectral Scenes," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2017, pp. 5108–5115.

[32] Y. Sun *et al.*, "RTFNet: RGB-Thermal Fusion Network for Semantic Segmentation of Urban Scenes," *IEEE Robotics and Automation Letters*, vol. 4, no. 3, pp. 2576–2583, 2019.

[33] Z. Liu *et al.*, "A ConvNet for the 2020s," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 11 976–11 986.

[34] A. Radford *et al.*, "Learning Transferable Visual Models From Natural Language Supervision," in *International Conference on Machine Learning (ICML)*. PMLR, 2021, pp. 8748–8763.

[35] T. Brown *et al.*, "Language Models are Few-Shot Learners," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, pp. 1877–1901, 2020.

[36] P. Xu *et al.*, "Multimodal Learning with Transformers: A Survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 10, pp. 12 113–12 132, 2023.

[37] J. Fu *et al.*, "Dual Attention Network for Scene Segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 3146–3154.

[38] J. Hu *et al.*, "Squeeze-and-Excitation Networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7132–7141.

[39] S. Mai *et al.*, "Analyzing Multimodal Sentiment Via Acoustic- and Visual-LSTM With Channel-Aware Temporal Convolution Network," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1424–1437, 2021.

[40] F. Chollet, "Xception: Deep Learning With Depthwise Separable Convolutions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1251–1258.

[41] 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.

[42] X. Zhu *et al.*, "Deformable DETR: Deformable Transformers for End-to-End Object Detection," in *International Conference on Learning Representations (ICLR)*, 2020.

[43] J. Jain *et al.*, "OneFormer: One Transformer to Rule Universal Image Segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 2989–2998.

[44] N. Carion *et al.*, "End-to-End Object Detection with Transformers," in *European Conference on Computer Vision (ECCV)*. Springer, 2020, pp. 213–229.

[45] A. Dosovitskiy *et al.*, "CARLA: An Open Urban Driving Simulator," in *Conference on Robot Learning (CoRL)*. PMLR, 2017, pp. 1–16.

[46] R. Fan *et al.*, "Road Surface 3D Reconstruction Based on Dense Subpixel Disparity Map Estimation," *IEEE Transactions on Image Processing*, vol. 27, no. 6, pp. 3025–3035, 2018.

[47] R. Fan *et al.*, "Rethinking Road Surface 3-D Reconstruction and Pothole Detection: From Perspective Transformation to Disparity Map Segmentation," *IEEE Transactions on Cybernetics*, vol. 52, no. 7, pp. 5799–5808, 2021.

[48] K. Perlin, "An Image Synthesizer," *ACM Siggraph Computer Graphics*, vol. 19, no. 3, pp. 287–296, 1985.

[49] L. Lipson *et al.*, "RAFT-Stereo: Multilevel Recurrent Field Transforms for Stereo Matching," in *2021 International Conference on 3D Vision (3DV)*. IEEE, 2021, pp. 218–227.

[50] M. Menze and 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.

[51] Y. Feng *et al.*, "D2NT: A high-performing depth-to-normal translator," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023, pp. 12 360–12 366.

[52] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization," in *International Conference on Learning Representations (ICLR)*, 2018.

[53] L. Chen *et al.*, "DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 834–848, 2017.

[54] H. Wang *et al.*, "Applying Surface Normal Information in Drivable Area and Road Anomaly Detection for Ground Mobile Robots," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 2706–2711.

[55] S. Gu *et al.*, "Histograms of the Normalized Inverse Depth and Line Scanning for Urban Road Detection," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 8, pp. 3070–3080, 2018.

[56] 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.

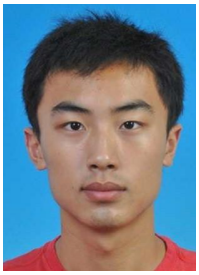
[57] L. Sun *et al.*, "Pseudo-LiDAR-Based Road Detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 8, pp. 5386–5398, 2022.

[58] A. A. Khan *et al.*, "LRDNet: Lightweight LiDAR Aided Cascaded Feature Pools for Free Road Space Detection," *IEEE Transactions on Multimedia*, 2022.

Authorized licensed use limited to: TONGJI UNIVERSITY. Downloaded on June 16,2024 at 05:25:01 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Intelligent Vehicles. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIV.2024.3388726

Jiahang Li (Graduate Student Member, IEEE) received his B.Eng. degree in automation from Taiyuan University of Technology in 2021. He is currently pursuing his Master's degree, under the supervision of Prof. Rui Fan, with the Machine Intelligence and Autonomous Systems (MIAS) Group in the Robotics and Artificial Intelligence Laboratory (RAIL) at Tongji University. His research interests include computer vision and deep learning.



Yikang Zhang obtained his B.Sc. degree in Automotion from the Beijing Institute of Technology in 2017, followed by an M.S. degree in ECE from UMASS Amherst in 2019, where he specialized in Model Predictive Control and Physical Unclonable Functions. He has since gained extensive industry experience, working with companies such as Tusimple, where he focused on truck state machine and controller, UnityDrive, where he contributed to Path Planner and chassis control, and CASIA, where he worked on Trajectory Prediction Algorithms. Currently, Yikang is pursuing his Ph.D. degree, supervised by Prof. Rui Fan. His research interests include Simulation, Planning, and Control in complex environment.

Peng Yun received the B.A. degree from Huazhong University of Science and Technology, Wuhan, China, in 2013, and the Ph.D. degree from HKUST, Hong Kong SAR, China, in 2017. He is currently working in DJI Automotive and conducting research on uncertainty modeling, life-long robotic learning, and perception algorithms.

Guangliang Zhou received his B.Sc. degree in Automation from Tongji University, Shanghai, China, in 2017, where he is currently pursuing his Ph.D. degree with the Robotics and Artificial Intelligence Laboratory. His research interests are in visual perception for robotics, with a focus on 6D object pose estimation and grasp detection.

Qijun Chen (Senior Member, IEEE) received the B.S. degree in automation from Huazhong University of Science and Technology, Wuhan, China, in 1987, the M.S. degree in information and control engineering from Xi'an Jiaotong University, Xi'an, China, in 1990, and the Ph.D. degree in control theory and control engineering from Tongji University, Shanghai, China, in 1999. He is currently a Full Professor in the College of Electronics and Information Engineering, Tongji University, Shanghai, China. His research interests include robotics control, environmental perception, and understanding of mobile robots and bioinspired control.

Rui Fan (Senior Member, IEEE) received the B.Eng. degree in Automation from the Harbin Institute of Technology in 2015 and the Ph.D. degree (supervisors: Prof. John G. Rarity and Prof. Naim Dahnoun) in Electrical and Electronic Engineering from the University of Bristol in 2018. He worked as a Research Associate (supervisor: Prof. Ming Liu) at the Hong Kong University of Science and Technology from 2018 to 2020 and a Postdoctoral Scholar-Employee (supervisors: Prof. Linda M. Zangwill and Prof. David J. Kriegman) at the University of California San Diego between 2020 and 2021. He began his faculty career as a Full Research Professor with the College of Electronics & Information Engineering at Tongji University in 2021, and was then promoted to a Full Professor in the same college, as well as at the Shanghai Research Institute for Intelligent Autonomous Systems in 2022.

Prof. Fan served as an associate editor for ICRA'23 and IROS'23/24, an area chair for ICIP'24, and a senior program committee member for AAAI'23/24. He is the general chair of the AVVision community and organized several impactful workshops and special sessions in conjunction with WACV'21, ICIP'21/22/23, ICCV'21, and ECCV'22. He was honored by being included in the Stanford University List of Top 2% Scientists Worldwide in both 2022 and 2023, recognized on the Forbes China List of 100 Outstanding Overseas Returnees in 2023, and acknowledged as one of Xiaomi Young Talents in 2023. His research interests include computer vision, deep learning, and robotics, with a specific focus on humanoid visual perception under the two-streams hypothesis.

Authorized licensed use limited to: TONGJI UNIVERSITY. Downloaded on June 16,2024 at 05:25:01 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.