RoadFormer+: Delivering RGB-X Scene Parsing Through Scale-Aware Information Decoupling and Advanced Heterogeneous Feature Fusion

Jianxin Huang, Jiahang Li, Ning Jia, Yuxiang Sun, *Member, IEEE*, Chengju Liu, Qijun Chen, *Senior Member, IEEE*, and Rui Fan, *Senior Member, IEEE*

***Abstract*—Task-specific data-fusion networks have marked considerable achievements in urban scene parsing. Among these networks, our recently proposed RoadFormer successfully extracts heterogeneous features from RGB images and surface normal maps and fuses these features through attention mechanisms, demonstrating compelling efficacy in RGB-Normal road scene parsing. However, its performance significantly deteriorates when handling other types/sources of data or performing more universal, all-category scene parsing tasks. To overcome these limitations, this study introduces RoadFormer+, an efficient, robust, and adaptable model capable of effectively fusing RGB-X data, where "X" represents additional types/modalities of data such as depth, thermal, surface normal, and polarization. Specifically, we propose a novel hybrid feature decoupling encoder to extract heterogeneous features and decouple them into global and local components. These decoupled features are then fused through a dual-branch multi-scale heterogeneous feature fusion block, which employs parallel Transformer attentions and convolutional neural network modules to merge multi-scale features across different scales and receptive fields. The fused features are subsequently fed into a decoder to generate the final semantic predictions. Notably, our proposed RoadFormer+ ranks first on the KITTI Road benchmark and achieves state-of-the-art performance in mean intersection over union on the Cityscapes, MFNet, FMB, and ZJU datasets. Moreover, it reduces the number of learnable parameters by 65% compared to RoadFormer. Our source code is publicly available at mias.group/RoadFormerPlus.**

***Index Terms*—Convolutional neural network, heterogeneous features, transformer, urban scene parsing.**

I. Introduction

SCENE parsing is crucial for the safety of autonomous driving [1]. With the widespread adoption of deep learning techniques, convolutional neural networks (CNNs) and Transformers have demonstrated significant performance improvements over traditional geometry-based models in various image segmentation tasks [2], [3], [4]. However, single-modal networks that rely solely on RGB images show limitations in handling challenging conditions such as poor illumination and adverse weather [5], [6]. To tackle these problems, subsequent research has explored the integration of useful information provided by additional data modalities. Depth or surface normal information has been utilized to identify spatially continuous regions [7], while thermal images have been employed to enhance object recognition robustness under poor lighting conditions [8]. Furthermore, polarization information has been used to improve segmentation performance for transparent and highly reflective objects [9]. Our recently proposed RoadFormer [1] effectively extracts heterogeneous features from RGB images and surface normal information and fuses these features for robust urban scene parsing, demonstrating notable efficacy in freespace and road defect detection. However, RoadFormer still has several limitations, especially when handling other types/sources of data. Moreover, the large quantity of parameters leads to considerable hardware resource consumption, thus limiting its deployment on terminal devices.

Most existing data-fusion networks use symmetric duplex encoders to extract heterogeneous features from multiple data sources and fuse them to provide a more comprehensive understanding of the environment [7], [10], [11]. However, while prior arts [1], [7], [12] have been developed to capture more discriminative features using these weight-separating duplex encoders, directly fusing these features may produce ambiguous features, thus negatively impairing the performance of scene parsing [13]. Additionally, the symmetric models with extensive parameters require more hardware resources for training, particularly when compared to networks that rely solely on RGB images [14]. Therefore, exploring an efficient and effective heterogeneous feature encoding strategy remains an under-explored research area that deserves more attention.

In addition to the heterogeneous feature extraction strategy, the performance of a data-fusion network also depends on the manner in which these features are fused. To address

Received 29 June 2024; revised 31 July 2024; accepted 16 August 2024. Date of publication 22 August 2024; date of current version 5 September 2025. This work was supported in part by the National Natural Science Foundation of China under Grant 62473288, Grant 62233013, Grant 62333017, Grant 62073245, Grant 62173248, and Grant 62403361, in part by the Science and Technology Commission of Shanghai Municipal under Grant 22511104500, in part by the Hong Kong Research Grants Council under Grant 15222523, and in part by the Fundamental Research Funds for the Central Universities, and Xiaomi Young Talents Program. *(Jianxin Huang and Jiahang Li are joint first authors.) (Corresponding author: Rui Fan.)*

Jianxin Huang, Jiahang Li, Ning Jia, Chengju Liu, Qijun Chen, and Rui Fan are with the College of Electronics and Information Engineering, Shanghai Research Institute for Intelligent Autonomous Systems, Shanghai Institute of Intelligent Science and Technology, the State Key Laboratory of Intelligent Autonomous Systems, and Frontiers Science Center for Intelligent Autonomous Systems, Tongji University, Shanghai 201804, China (e-mail: jasonhuang@tongji.edu.cn; lijiahang617@tongji.edu.cn; jianing7072@tongji.edu.cn; liuchengju@tongji.edu.cn; qjchen@tongji.edu.cn; rui.fan@ieee.org).

Yuxiang Sun is with the Department of Mechanical Engineering, City University of Hong Kong, Kowloon, Hong Kong (e-mail: yx.sun@cityu.edu.hk).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TIV.2024.3448251.

Digital Object Identifier 10.1109/TIV.2024.3448251

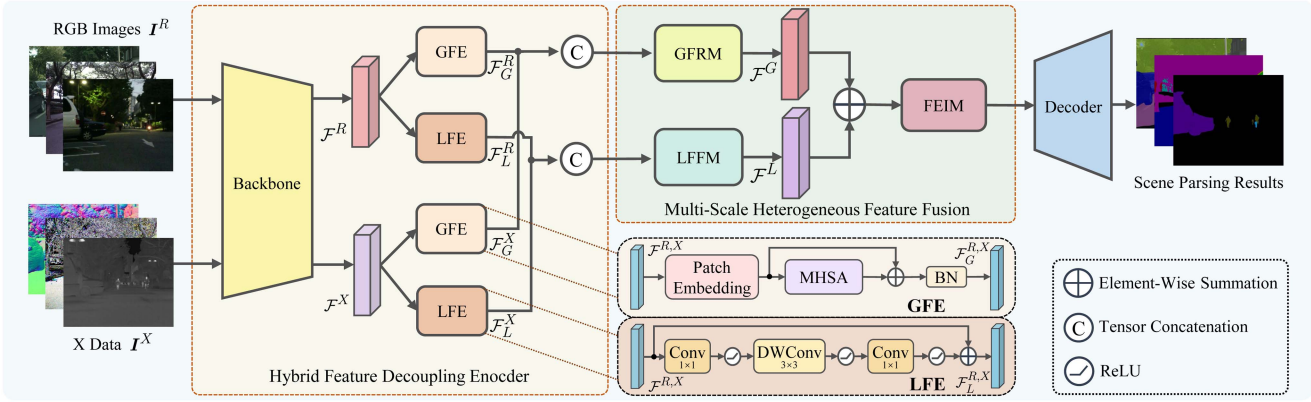

Fig. 1. An overview of our proposed RoadFormer+ architecture.

this issue, recent works [1], [12], [14], [15] employ learnable feature fusion approaches, which significantly outperform traditional, non-discriminative fusion methods that rely solely on element-wise concatenation or summation [5], [7]. For example, RoadFormer [1] adopts a Transformer-based approach to effectively capture long-range dependencies within heterogeneous features. On the other hand, RDFNet [16] employs CNN-based modules to process multi-scale features, effectively extracting local spatial cues, such as edges and textures, within a relatively small receptive field. However, these methods typically employ single-branch feature fusion blocks, where features extracted from RGB images and additional data types (referred to as "X" data) are fused using convolutional layers or attention mechanisms. Such single-branch feature fusion strategies may not always effectively encode both local and global contexts simultaneously, limiting their capacity to fully exploit the advantages of multi-modal/source data fusion. Considering Transformers' remarkable capability in modeling long-range dependencies and CNNs' robustness in local feature extraction [13], further research into combining capabilities of CNNs' local feature integration and Transformer's global representation modeling through a dual-branch fusion design to enhance scene parsing is highly warranted.

Moreover, while task-specific networks such as RoadFormer demonstrate impressive performance in RGB-Normal road scene parsing, their applicability to more universal urban scene parsing tasks and their effectiveness in handling diverse data types remain limited. For instance, RoadFormer exhibits a significant performance drop on comprehensive scene parsing datasets, such as the KITTI Semantics [17] and Cityscapes [18], compared to existing state-of-the-art (SoTA) RGB-D/Normal methods. Additionally, it performs suboptimally when processing RGB-Thermal/Polarization data [8], [9]. It is urged to design a universal RGB-X data-fusion network that performs robustly across multiple data sources for urban scene parsing.

To address the aforementioned limitations, we first design a more efficient hybrid feature decoupling encoder (HFDE). Given the correlation between RGB images and their corresponding X data, we first replace the duplex encoder with a weight-sharing backbone to reduce the number of learnable parameters. We then employ an asymmetric architecture that independently utilizes two global feature enhancers (GFEs) and two local feature extractors (LFEs) to decouple heterogeneous features, effectively modeling their inherent differences at various scales. Subsequently, we introduce a robust dual-branch multi-scale heterogeneous feature fusion (MHFF) block to fuse heterogeneous features in parallel, ensuring a comprehensive integration of global and local features. The MHFF block utilizes Transformer-based and CNN-based modules to parallelly fuse and calibrate multi-scale features. Our proposed RoadFormer+, as illustrated in Fig. 1, an upgraded version of RoadFormer, with all these innovative components incorporated, demonstrates superior performance over RoadFormer across four RGB-Normal scene parsing datasets, while reducing the learnable parameters by around 65%. Furthermore, RoadFormer+ achieves SoTA performance in RGB-Normal, RGB-Thermal, and RGB-Polarization scene parsing, demonstrating its exceptional applicability across a broad range of RGB-X data-fusion scenarios.

Our contributions can be summarized as follows:

- We introduce HFDE, which consists of a weight-sharing backbone and two pairs of independent GFEs and LFEs, to extract heterogeneous features and effectively capture both the correlation and inherent differences between RGB images and X data.
- We design a dual-branch MHFF block to capture both global and local features simultaneously. It seamlessly integrates Transformer-based and CNN-based modules, so as to utilize different receptive fields to achieve advanced heterogeneous feature fusion.
- We propose RoadFormer+, a novel urban scene parsing approach with fewer parameters compared with RoadFormer, which achieves SoTA performance across multiple RGB-X scene parsing datasets.

The remainder of this article is organized as follows: In Section II, we review related works on urban scene parsing. In Section III, we introduce our proposed RoadFormer+. In Section IV, we present quantitative and qualitative experimental results and their corresponding analyses. Finally, in Section V, we conclude this work and discuss potential future directions.

II. Related Work

A. Single-Modal Scene Parsing

Since the introduction of FCN [19], various CNN-based scene parsing networks have been developed. For instance, PSPNet [20] uses pyramid pooling to capture semantic information at multiple scales. DeepLabV3+ [21] employs atrous convolutions

with different dilation rates to enrich the contextual feature encoding across scales. Additionally, MobileNetV2 [22] adopts lighter architectures based on depth-wise separable convolutions to reduce model parameters and computational demands. In these CNN-based networks, each convolutional kernel processes only a local region of the image at a time. This local receptive field design enables CNNs to excel at extracting local features, such as edges and textures [23].

Transformers have gained prominence in scene parsing tasks due to their exceptional global aggregation capabilities compared to CNNs [24]. The attention mechanisms within Transformers allow each token to interact with all others simultaneously [25]. These interactions help achieve a comprehensive understanding of the correlation between each token and the global context, thereby better extracting global features. Segmentation Transformer (SETR) [26], pioneering the use of a Transformer-based architecture for scene parsing, adopts a method similar to the vision Transformer (ViT) [27] by tokenizing images into patches and processing them through Transformer blocks to enhance the global context modeling in the encoder. Furthermore, the MaskFormer series [28], [29] introduces a novel Transformer-based decoding paradigm by segmenting images into a set of masks, each associated with a class prediction. This mask classification paradigm, previously validated in [1], has been effectively incorporated into our enhanced RoadFormer+ design, further optimizing its performance.

B. Data-Fusion Scene Parsing

Scene parsing networks that rely solely on RGB images have been found to be highly sensitive to environmental factors such as lighting and weather conditions [7]. To overcome this limitation, data-fusion networks effectively utilize heterogeneous features extracted from RGB images and additional data sources. FuseNet [5] pioneered the incorporation of depth information into scene parsing. It uses independent CNN encoders for RGB and depth images and fuses their features through element-wise summation. MFNet [8] and RTFNet [30] strike a balance between speed and accuracy in RGB-Thermal driving scene parsing. Inspired by [5], the SNE-RoadSeg series [7], [31] incorporates surface normal information into freespace detection. These networks employ densely connected skip connections to enhance feature decoding. Despite the improved performance achieved by these networks, the simplistic feature fusion strategies potentially restrict their capacity to fully exploit the complementary information present in heterogeneous features.

To address this challenge, recent studies have employed more advanced and learnable feature fusion strategies. RoadFormer [1] combines self-attention with channel attention to form a novel feature synergy block that greatly enhances the fusion of heterogeneous features. Data-fusion networks have also garnered attention in the broader domain of scene parsing. Recent works CMX [12] and CAINet [14] utilize various attention modules to effectively fuse and recalibrate heterogeneous features. Additionally, SASEM [32] introduces a plug-and-play module to enhance semantic supervision, thereby improving feature recovery capabilities. Moreover, CDDFuse [13] implements a two-step training strategy that integrates CNN and Transformer blocks in parallel to fuse multi-modal medical images effectively. This article delves into more robust and general-purpose modules so as to more effectively fuse heterogeneous features. Our proposed RoadFormer+ not only broadens its applicability and generalizability to a wider range of scene parsing tasks but also significantly reduces the number of model parameters.

III. Methodology

A. Hybrid Feature Decoupling Encoder

1) Overall Feature Encoding Pipeline: Current networks generally employ symmetric duplex encoders to extract heterogeneous features from multiple data sources [1], [7]. However, such dual-branch designs not only double the number of learnable parameters in the feature encoding phase but may also potentially lead to feature conflict [33]. To address this issue, we propose an HFDE to improve the efficiency of heterogeneous feature extraction. Specifically, considering the correlation between heterogeneous features [34], we first employ a weight-sharing backbone to process the given RGB image $\boldsymbol{I}^R \in \mathbb{R}^{H\times W\times 3}$ and its corresponding X data $\boldsymbol{I}^X \in \mathbb{R}^{H\times W\times 3}$, thereby generating multi-scale features $\mathcal{F}^R = \{\boldsymbol{F}_1^R, \ldots, \boldsymbol{F}_4^R\}$ and $\mathcal{F}^X = \{\boldsymbol{F}_1^X, \ldots, \boldsymbol{F}_4^X\}$. For X data with a single channel (such as depth, thermal, and polarization information), we replicate it three times along the channel dimension to match the RGB image's dimensions of $H \times W \times 3$. Here, $\boldsymbol{F}_i^{R,X} \in \mathbb{R}^{\frac{H}{S_i}\times\frac{W}{S_i}\times C_i}$ represents the features in the i-th encoding stage, where C_i and $S_i = 2^{i+1}$ $(i \in [1,4] \cap \mathbb{Z})$ denote the channel and stride numbers, respectively, and H and W denote the height and width of the input image, respectively. Furthermore, we employ two weight-separating GFEs and LFEs to extract global features $\mathcal{F}_G^{R,X}$ and local features $\mathcal{F}_L^{R,X}$ at four spatial scales from the heterogeneous features $\mathcal{F}^{R,X}$, respectively, thereby realizing feature decoupling. Finally, $\mathcal{F}_G^{R,X}$ and $\mathcal{F}_L^{R,X}$ are fed into the MHFF block for further feature fusion and recalibration.

2) Weight-Sharing Backbone: Large-kernel convolutions exhibit considerable potential in capturing long-range dependencies, owing to their expansive receptive fields, while still retaining favorable inductive biases crucial for vision-specific tasks such as scene parsing [35]. For instance, the areas surrounding vehicles are more likely to be roads rather than buildings. In our previous study [1], ConvNeXt [36] demonstrates superior performance compared to ResNet [37] and Swin Transformer [38], and thus, we continue to adopt it as the backbone in this study. We also compare the performance of ConvNeXt with the recently proposed SoTA backbone networks UniRepLKNet [35] and DiNAT [39]. Detailed experimental results and analyses are provided in Table X. Our backbone is constructed using two identical, weight-sharing ConvNeXt models.

3) Global Feature Enhancer: ViT has shown exceptional performance across various fundamental vision tasks [27], [38]. Its self-attention mechanism effectively models the global receptive field, thereby enhancing the contextual understanding essential for recognizing large continuous areas such as roads and sidewalks. Consequently, we utilize a GFE based on the multi-head self-attention mechanism to further emphasize the long-range global features. Given the robustness of the backbone network, we omit the positional encoding and replace the commonly used feed-forward network layers with simple normalization operations to reduce the number of model parameters. RGB features $\mathcal{F}^R$ and X features $\mathcal{F}^X$ are respectively mapped to query, value, and key matrices through convolutional layers. We also introduce a residual connection into the attention operation.

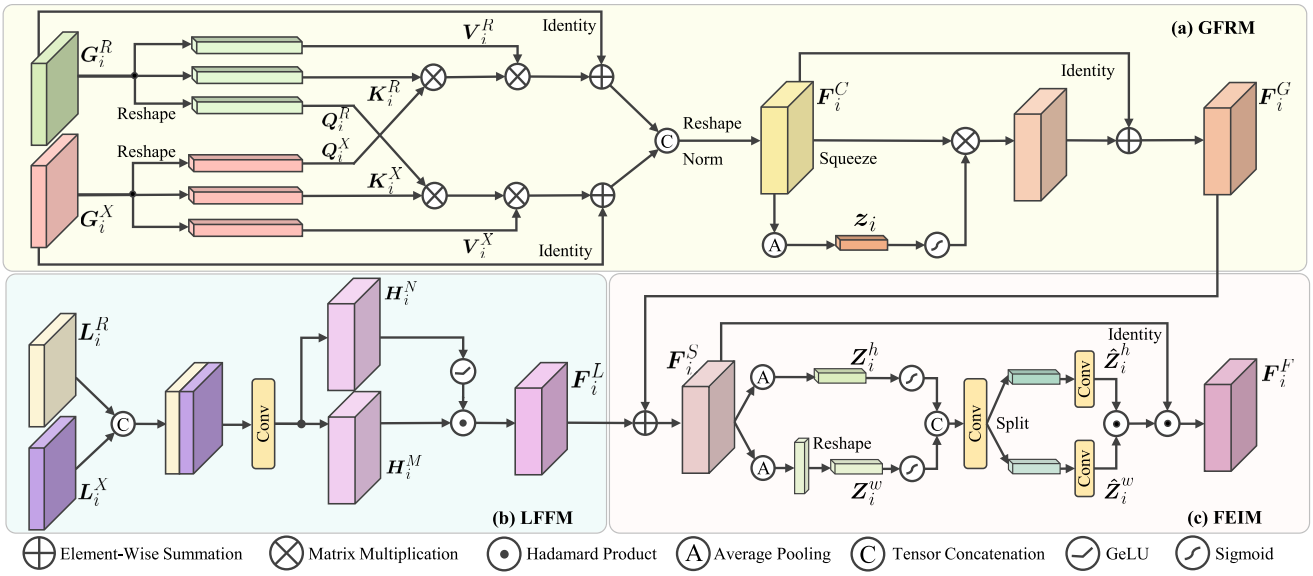

Fig. 2. An illustration of our proposed multi-scale heterogeneous feature fusion block, consisting of (a) a global feature recalibration module, (b) a local feature fusion module, and (c) a feature enhancement and integration module.

Our GFE module can be formulated as follows:

$$\boldsymbol{G}_i = \text{Norm}\left(\text{MHSA}(\boldsymbol{F}_i) + \boldsymbol{F}_i\right), \tag{1}$$

where $\boldsymbol{F}_i$ represents the i-th feature maps within $\mathcal{F}^R$ and $\mathcal{F}^X$, $\boldsymbol{G}_i$ represents the i-th feature maps within $\mathcal{F}_G^R$ and $\mathcal{F}_G^X$, and MHSA represents the multi-head self-attention mechanism operation. After processing by the GFE, the enhanced global features $\mathcal{F}_G^R$ and $\mathcal{F}_G^X$ are obtained.

4) Local Feature Extractor: Local detail features, such as edges and corners, are crucial for accurate scene parsing. Compared to Transformers, convolution operations are proficient at extracting local features and further enhancing fine-grained details [40]. Therefore, we propose a LFE, which incorporates an inverted residual block from MobileNetV2 [22] to process $\mathcal{F}^R$ and $\mathcal{F}^X$, specifically targeting local features. This lightweight module strikes a balance between model parameters and accuracy, as demonstrated across multiple tasks [14]. Our LFE can be formalized as follows:

$$\boldsymbol{L}_i = \underset{1\times1}{\text{Conv}}\left(\text{ReLU}\left(\underset{3\times3}{\text{DWConv}}\left(\text{ReLU}\left(\underset{1\times1}{\text{Conv}}(\boldsymbol{F}_i)\right)\right)\right)\right) + \boldsymbol{F}_i, \tag{2}$$

where $\boldsymbol{L}_i$ denotes the i-th feature maps within $\mathcal{F}_L^R$ and $\mathcal{F}_L^X$. After processing by the LFE, we obtain the local features $\mathcal{F}_L^R$ and $\mathcal{F}_L^X$.

B. Multi-Scale Heterogeneous Feature Fusion Block

To further fuse and integrate global and local features, we introduce a dual-branch MHFF block, which employs attention mechanisms and CNN modules in parallel. An MHFF consists of (1) a global feature recalibration module (GFRM) that utilizes a cross-attention mechanism to recalibrate $\mathcal{F}_G^R$ and $\mathcal{F}_G^X$, (2) a local feature fusion module (LFFM) that utilizes convolutional layers to fuse $\mathcal{F}_L^R$ and $\mathcal{F}_L^X$, and (3) a feature enhancement and integration module (FEIM) based on a spatial attention mechanism to integrate heterogeneous features and generate robust fused feature $\mathcal{F}^F$.

1) Global Feature Recalibration Module: Heterogeneous global features $\mathcal{F}_G^R$ and $\mathcal{F}_G^X$ are generally complementary [34]. For example, road areas often appear consistent in color across RGB images and possess uniform normal vectors and polarization properties. Therefore, one feature type can be utilized to mitigate potential noise in its complementary feature type [12]. Additionally, features from different channels do not all contribute positively to semantic predictions [1], [41], necessitating the recalibration of heterogeneous features along the channel dimension [42]. To address these challenges, we introduce the GFRM (see Fig. 2(a)) to recalibrate and fuse $\mathcal{F}_G^R$ and $\mathcal{F}_G^X$. The cross-attention mechanism, which considers interactions among all positions in the input, is well-suited for calibrating complementary heterogeneous global features and has demonstrated excellent performance in many visual tasks [12]. Drawing inspiration from these approaches, the GFRM first recalibrates global features using a cross-attention mechanism, which can be formulated as follows:

$$\boldsymbol{G}_i^{R'} = \text{Softmax}\left(\boldsymbol{Q}_i^R \boldsymbol{K}_i^{X\top}\right)\kappa_i \boldsymbol{V}_i^X + \boldsymbol{G}_i^R, \tag{3}$$

$$\boldsymbol{G}_i^{X'} = \text{Softmax}\left(\boldsymbol{Q}_i^X \boldsymbol{K}_i^{R\top}\right)\gamma_i \boldsymbol{V}_i^R + \boldsymbol{G}_i^X, \tag{4}$$

$$\boldsymbol{F}_i^C = \text{Norm}\left(\delta\left([\boldsymbol{G}_i^{R'}, \boldsymbol{G}_i^{X'}]\right)\right), \tag{5}$$

where $\boldsymbol{G}_i^R$ and $\boldsymbol{G}_i^X$ represent the i-th feature maps within $\mathcal{F}_G^R$ and $\mathcal{F}_G^X$, respectively, $\boldsymbol{G}_i^R$ and $\boldsymbol{G}_i^X$ are then identically mapped to query $\boldsymbol{Q}_i^{R,X}$, key $\boldsymbol{K}_i^{R,X}$ and value $\boldsymbol{V}_i^{R,X}$ embeddings, $[\cdot,\cdot]$ denotes the concatenation operation along the channel dimension, and δ is a non-linear activation function. The learnable coefficients κ_i and γ_i can adaptively adjust attention significance [43]. For $\mathcal{F}^C$, we further employ channel-wise attention to emphasize key features and suppress those with low information

density, which can be formulated as follows:

$$z_{i,j} = \frac{S_i^2}{HW} \sum_{h=1}^{\frac{H}{S_i}} \sum_{w=1}^{\frac{W}{S_i}} \boldsymbol{F}_i^C(h, w, j), \tag{6}$$

$$\boldsymbol{F}_i^G = \boldsymbol{F}_i^C \odot \sigma\left(\underset{1\times1}{\text{Conv}}(\boldsymbol{z}_i)\right) + \boldsymbol{F}_i^C, \tag{7}$$

where $\boldsymbol{z}_i = [z_{i,1}, \ldots, z_{i,C_i}] \in \mathbb{R}^{1\times1\times C_i}$ stores the average pooling results of each feature map in $\boldsymbol{F}_i^C$, σ is the sigmoid function, and $\odot$ denotes the Hadamard product operation. Finally, we obtain the fused global feature $\mathcal{F}^G = \{\boldsymbol{F}_1^G, \ldots, \boldsymbol{F}_4^G\}$. Here, $\boldsymbol{F}_i^G \in \mathbb{R}^{\frac{H}{S_i}\times\frac{W}{S_i}\times C_i}$ represents the global features in the i-th feature fusion stages.

2) Local Feature Fusion Module: To preserve more local contexts when fusing heterogeneous features $\mathcal{F}_L^R$ and $\mathcal{F}_L^X$, we propose a convolution-based LFFM (see Fig. 2(b)). Inspired by the MLP-Mixer [44], our LFFM captures relationships between heterogeneous features from different local regions, generating fused local features. The LFFM can be mathematically represented as follows:

$$\boldsymbol{H}_i^L = \underset{3\times3}{\text{DWConv}}\left(\text{Conv}([\boldsymbol{L}_i^R, \boldsymbol{L}_i^X])\right), \tag{8}$$

where $\boldsymbol{L}_i^R$ and $\boldsymbol{L}_i^X$ represent the i-th feature maps within $\mathcal{F}_L^R$ and $\mathcal{F}_L^X$, respectively, each having C_i channels. After concatenating $\boldsymbol{L}_i^R$ and $\boldsymbol{L}_i^X$ along the channel dimension, we employ depth-wise separable convolutions to expand their channels to $4C_i$, thereby enhancing the local context. The resultant $\boldsymbol{H}_i^L$ is then split into $\boldsymbol{H}_i^M$ and $\boldsymbol{H}_i^N$ along the channel dimension. This design allows the model to learn new feature representations, which is further validated in Table XI. These two components interact through Hadamard multiplication, enabling the capture of relationships between features from different local regions:

$$\boldsymbol{F}_i^L = \underset{1\times1}{\text{Conv}}\left(\boldsymbol{H}_i^M \odot \sigma(\boldsymbol{H}_i^N)\right), \tag{9}$$

where we utilize the Gaussian error linear unit (GELU) as the activation function $\sigma(\cdot)$. Then we obtain the fused local features $\mathcal{F}^L = \{\boldsymbol{F}_1^L, \ldots, \boldsymbol{F}_4^L\}$, where $\boldsymbol{F}_i^L \in \mathbb{R}^{\frac{H}{S_i}\times\frac{W}{S_i}\times C_i}$ represents the local features in the i-th feature fusion stage.

3) Feature Enhancement and Integration Module: Spatial information is crucial for capturing spatial structures in visual perception tasks [15]. Nonetheless, our GFRM and LFFM fuse heterogeneous features across channel dimensions, squeezing spatial information into a channel descriptor, and hence is difficult to preserve positional information [45]. Therefore, it is necessary to introduce additional spatial information when integrating global features $\mathcal{F}^G$ and local features $\mathcal{F}^L$. The spatial attention mechanism emphasizes the importance of specific regions within features, aiding the network in focusing on "where" informative parts are located [46]. Inspired by the coordinated attention [45], we introduce the FEIM (see Fig. 2(c)) to further enhance and integrate $\mathcal{F}^G$ and $\mathcal{F}^L$, enabling the module to detect more subtle spatial variations. Specifically, we employ global pooling kernels $(H, 1)$ or $(1, W)$ to aggregate features along the height and width dimensions, respectively. Thus, the output of the j-th channel at height p and width q can be formulated as:

$$\boldsymbol{F}_i^S = \boldsymbol{F}_i^G + \boldsymbol{F}_i^L, \tag{10}$$

$$z_{i,j,p}^h = \frac{1}{W} \sum_{0\le m<W} \boldsymbol{F}_i^S(p, m, j), \tag{11}$$

$$z_{i,j,q}^w = \frac{1}{H} \sum_{0\le n<H} \boldsymbol{F}_i^S(n, q, j), \tag{12}$$

where $\boldsymbol{F}_i^G$ and $\boldsymbol{F}_i^L$ represent the i-th feature maps within $\mathcal{F}^G$ and $\mathcal{F}^L$, respectively, and $\boldsymbol{Z}_i^h \in \mathbb{R}^{H\times1\times C_i}$ as well as $\boldsymbol{Z}_i^w \in \mathbb{R}^{1\times W\times C_i}$ store the average pooling results of each feature map in $\boldsymbol{F}_i^S$ across the dimensions of H and W, respectively. $\boldsymbol{Z}_i^w$ is subsequently reshaped into $\mathbb{R}^{W\times1\times C_i}$. We further apply a convolutional layer and a Sigmoid function to make full use of the captured positional information, enhancing the network's ability to accurately emphasize regions of interest. This process can be formulated as follows:

$$\hat{\boldsymbol{Z}}_i = \sigma\left(\underset{1\times1}{\text{Conv}}([\boldsymbol{Z}_i^h, \boldsymbol{Z}_i^w])\right), \tag{13}$$

where $[\cdot,\cdot]$ denotes the concatenation operation along the spatial dimension. Then, $\hat{\boldsymbol{Z}}_i$ is split into two separate tensors: $\hat{\boldsymbol{Z}}_i^h \in \mathbb{R}^{H\times1\times C_i}$ and $\hat{\boldsymbol{Z}}_i^w \in \mathbb{R}^{W\times1\times C_i}$. This allows interactions between $\hat{\boldsymbol{Z}}_i^h$ and $\hat{\boldsymbol{Z}}_i^w$ from both dimensions, enhancing the emphasis on regions of interest. $\hat{\boldsymbol{Z}}_i^w$ is then reshaped into $\mathbb{R}^{1\times W\times C_i}$. Each element within the two attention maps, $\hat{\boldsymbol{Z}}_i^h$ and $\hat{\boldsymbol{Z}}_i^w$, indicates the presence of objects of interest across respective rows or columns. $\hat{\boldsymbol{Z}}_i^h$ and $\hat{\boldsymbol{Z}}_i^w$ are applied to $\boldsymbol{F}_i^S$ to more accurately pinpoint the exact location of the object of interest, which can be written as follows:

$$\boldsymbol{F}_i^F = \boldsymbol{F}_i^S \odot \hat{\boldsymbol{Z}}_i^h \odot \hat{\boldsymbol{Z}}_i^w. \tag{14}$$

Finally, we obtain the fused features $\mathcal{F}^F = \{\boldsymbol{F}_1^F, \ldots, \boldsymbol{F}_4^F\}$, which are forwarded to the decoder to obtain the final semantic prediction. Given the outstanding performance of the multi-scale Transformer decoder employed in RoadFormer, we retain this design. Please refer to [1] for more details on the decoder and loss function.

IV. Experiments

A. Datasets

We compare RoadFormer+ with other SoTA scene parsing networks on the following seven RGB-X datasets:

1) SYN-UDTIRI [1]: This dataset contains over 10,000 pairs of stereo road images, along with corresponding depth maps, surface normal information, and semantic annotations, including three categories: freespace, road defect, and other objects. It is created using the CARLA simulator [47] and first introduced in our previous work [1]. The input images are resized to a resolution of 640 × 352 pixels.

2) KITTI Road [48]: This dataset has 289 pairs of stereo road images and their corresponding LiDAR point clouds for both model training and validation. We employ a data pre-processing strategy akin to that detailed in [7]. The input images are resized to a resolution of 1,280 × 384 pixels.

3) Cityscapes [18]: This widely used urban scene dataset contains 2,975 stereo training images and 500 validation images, with well-annotated semantic annotations. Notably, the surface normal information is derived from depth images generated

TABLE I
QUANTITATIVE COMPARISON OF ROAD DEFECT DETECTION ON THE SYN-UDTIRI TEST SET

Method	Publication	IoU (%) ↑	Fsc (%) ↑	Pre (%) ↑	Rec (%) ↑	#Params (M) ↓
OFF-Net	ICRA'22 [3]	83.80	91.20	91.90	90.50	25.2
RTFNet	RAL'19 [30]	90.50	95.00	95.50	94.50	254.5
DFormer	ICLR'24 [53]	90.88	95.22	96.09	94.37	38.8
CAINet	T-MM'24 [14]	91.77	95.71	95.43	95.99	**12.2**
SNE-RoadSeg	ECCV'20 [7]	92.10	95.90	96.70	95.10	201.3
CMX	T-ITS'23 [12]	93.31	96.27	96.54	96.81	138.8
RoadFormer (B)	T-IV'24 [1]	93.06	96.41	96.19	96.63	206.8
RoadFormer (L)	T-IV'24 [1]	93.51	96.65	96.61	96.69	438.6
RoadFormer+ (B)	Ours	**94.11**	**96.96**	**97.03**	**96.90**	152.4

using RAFT-Stereo [49], trained on the KITTI dataset [50]. The input images are resized to a resolution of 1,024 × 512 pixels.

4) KITTI Semantics [17]: This dataset contains 200 images and their corresponding semantic annotations across 19 classes. Surface normal information is derived from depth data acquired using ViTAStereo [51], chosen for its superior accuracy. The input images are resized to a resolution of 1,280 × 384 pixels.

5) MFNet [8]: This urban driving scene parsing dataset contains 1,569 synchronized pairs of RGB and thermal images at a resolution of 640 × 480 pixels. It includes semantic annotations across nine classes: bike, person, car, road lanes, guardrail, car stop, bump, color cone, and background.

6) FMB [52]: This dataset contains 1,500 well-rectified RGB-Thermal image pairs (resolution: 800 × 600 pixels), captured in urban driving scenes under different illumination conditions. It provides semantic annotations across 14 classes.

7) ZJU [9]: This RGB-Polarization dataset, designed for automated driving applications, contains 394 image pairs. Each pair contains four polarized images captured at different polarization angles (0°, 45°, 90°, and 135°). The input images are resized to a resolution of 612 × 512 pixels.

B. Experimental Setup and Evaluation Metrics

For the SYN-UDTIRI and other RGB-Normal datasets, we exclusively use surface normal information estimated using the D2NT algorithm [54] owing to its superior accuracy. This information serves as the "X" data to train our RoadFormer+. Additionally, depth, thermal, and polarization information are replicated across the channel dimension three times during data pre-processing to match the $H \times W \times 3$ dimensions of RGB images. During training, we employ the common data augmentation techniques used in semantic segmentation, including resizing, random cropping, and random flipping of RGB-X image pairs. Additionally, we make random adjustments to the brightness, contrast, saturation, and hue of the RGB images. All networks are trained for the same number of epochs on an NVIDIA RTX 3090 GPU using the AdamW optimizer [55], with a polynomial decay strategy for the learning rate [23]. The initial learning rate is set to 10^{-4} with a weight decay of 5×10^{-2}, and learning rate multipliers of 10^{-1} are applied to the weight-sharing backbone.

We evaluate the performance of our models using five common metrics: accuracy (Acc), precision (Pre), recall (Rec), 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 and KITTI Semantics benchmarks are available on the official webpage: cvlibs.net/datasets/kitti.

TABLE II
COMPARISON WITH SOTA ALGORITHMS PUBLISHED ON THE KITTI ROAD BENCHMARK

Method	MaxF (%) ↑	Pre (%) ↑	Rec (%) ↑	Rank
SNE-RoadSeg [7]	96.75	96.90	96.61	13
RoadFormer (B) [1]	97.50	97.16	**97.84**	3
SNE-RoadSegV2 [31]	97.55	**97.57**	97.53	2
RoadFormer+ (B)	**97.56**	97.43	97.69	**1**

TABLE III
QUANTITATIVE COMPARISON OF FREESPACE DETECTION ON THE VALIDATION SET OF THE CITYSCAPES DATASET

Method	IoU (%) ↑	Fsc (%) ↑	Acc (%) ↑
SNE-RoadSeg [7]	93.22	96.49	97.68
SNE-RoadSegV2 [31]	94.40	97.12	98.11
RoadFormer (B) [1]	95.87	97.89	**98.30**
RoadFormer+ (B)	**96.01**	**97.96**	97.82

TABLE IV
QUANTITATIVE COMPARISON OF ALL-CATEGORY SCENE PARSING ON THE VALIDATION SET OF THE CITYSCAPES DATASET

	Method	mIoU (%) ↑	mFsc (%) ↑	mAcc (%) ↑
RGB	SegFormer [56]	64.51	76.99	76.39
	DeepLabV3+ [23]	68.66	80.34	78.89
	ConvNeXt [36]	73.35	83.94	83.32
	Mask2Former [29]	74.78	84.97	85.90
RGB-Depth	CAINet [14]	62.38	75.04	73.68
	CMX [12]	74.11	84.41	83.30
	DFormer [53]	74.37	84.55	84.00
	RoadFormer (B) [1]	76.09	85.83	86.30
	RoadFormer+ (B)	77.42	86.72	86.23
RGB-Normal	RTFNet [30]	49.60	61.20	**90.00**
	SNE-RoadSeg [7]	53.40	64.54	85.64
	CAINet [14]	62.41	75.13	74.23
	CMX [12]	73.50	83.99	83.67
	RoadFormer (B) [1]	76.18	85.88	85.38
	RoadFormer+ (B)	77.57	86.84	86.77
	RoadFormer+ (L)	**78.53**	**87.48**	87.00

C. Comparison With SoTA Networks

We first conduct experiments on four RGB-Normal datasets. The quantitative results on the SYN-UDTIRI, Cityscapes, KITTI Road, and KITTI Semantics datasets are presented in Tables I–V, respectively. In these experiments, the symbols "B" and "L" respectively denote the use of ConvNeXt-B and ConvNeXt-L as the backbones. These results demonstrate that our proposed RoadFormer+ significantly outperforms all other SoTA networks, including our previous work RoadFormer [1],

TABLE V
COMPARISON WITH SOTA ALGORITHMS PUBLISHED ON THE KITTI SEMANTICS BENCHMARK

Method	IoU Class (%) ↑	IoU Category (%) ↑	Rank
RoadFormer (B) [1]	67.17	87.89	5
VideoProp-LabelRelax [57]	72.82	**88.99**	4
RoadFormer+ (B)	70.32	87.16	-
RoadFormer+ (L)	**73.13**	88.75	**3**

TABLE VI
QUANTITATIVE COMPARISON ON THE MFNET TEST SET

Method	mIoU (%) ↑	Rank
RTFNet [30]	53.2	33
RoadFormer (B) [1]	58.0	12
CAINet [14]	58.6	9
CMX [12]	59.7	5
CMNeXt [58]	59.9	4
CRM-RGBTSeg [59]	61.4	3
HAPNet [60]	61.5	2
RoadFormer+ (B)	60.9	-
RoadFormer+ (L)	**62.7**	**1**

TABLE VII
QUANTITATIVE COMPARISON ON THE FMB DATASET

Method	mIoU (%) ↑	Rank
SegMiF [52]	54.8	4
RoadFormer (B) [1]	69.2	2
RoadFormer+ (B)	73.1	-
RoadFormer+ (L)	**74.1**	**1**

TABLE VIII
QUANTITATIVE COMPARISON ON THE ZJU-RGB-P DATASET

Method	mIoU (%) ↑	Rank
EAFNet [9]	85.7	5
RoadFormer (B) [1]	92.6	4
CMX [12]	92.6	3
RoadFormer+ (B)	92.9	-
RoadFormer+ (L)	**93.0**	**1**

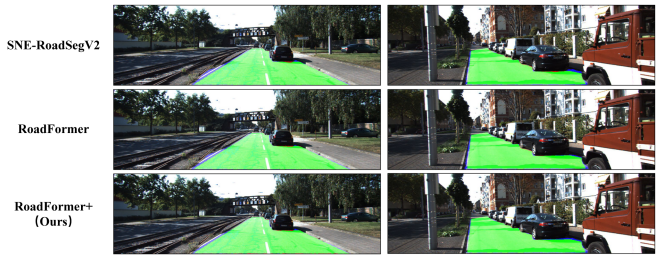

Fig. 3. Qualitative comparison between our proposed RoadFormer+ and other SoTA networks on the KITTI Road dataset. The results are produced by the official KITTI online benchmark suite. The classifications are visualized with true positives in green, false positives in blue, and false negatives in red.

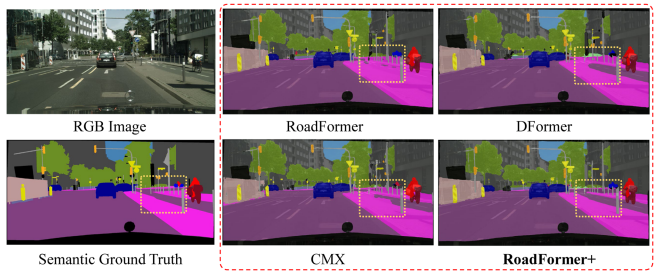

Fig. 4. Qualitative comparisons between our proposed RoadFormer+ and other SoTA networks on the Cityscapes validation set, where significantly improved regions are shown with yellow dashed boxes.

across all four RGB-Normal datasets. This validates its exceptional performance and robustness in effectively parsing various types of road scenes. Notably, as shown in Table I, RoadFormer+ based on ConvNeXt-B reduces the number of learnable parameters by 65% compared to RoadFormer.

Furthermore, we conduct experiments on the Cityscapes dataset by treating it as both a binary segmentation task (road versus background) and a full-category segmentation task (19 labeled categories plus an "ignore" category). Experimental results are presented in Tables III and IV, respectively. We also compare RoadFormer+ with four SoTA single-modal networks. It is worth noting that traditional data-fusion networks, which typically employ basic element-wise addition or feature-level concatenation for feature fusion, perform worse than single-modal networks. This underperformance may be attributed to the noise present in disparity maps used for surface normal estimation, which are derived directly from a stereo matching network pre-trained on the KITTI dataset. Experimental results further demonstrate that RoadFormer+ effectively overcomes this issue through feature recalibration and enhancement, thus preventing performance degradation even when surface normal information is inaccurate.

We submit the test set results obtained by RoadFormer+ to both the KITTI Road and KITTI Semantics benchmarks for performance comparison. As shown in Tables II and V, RoadFormer+ ranks first on the KITTI Road benchmark and ranks third on the KITTI Semantics benchmark. Notably, the top-performing SoTA methods in the KITTI Semantics benchmark employ sequential frames (± 10) from the scene flow subset for data augmentation. Despite this, RoadFormer+ exhibits superior performance in urban scene parsing compared to all previously published methods.

Furthermore, we explore the applicability of RoadFormer+ for RGB-Thermal and RGB-Polarization scene parsing. Experimental results on three public datasets, MFNet (RGB-Thermal), FMB (RGB-Thermal), and ZJU (RGB-Polarization), demonstrate the superiority of RoadFormer+ over other task-specific data-fusion networks for these modalities. Impressively, RoadFormer+ achieves an improvement in mIoU of 1.2–9.5% on the MFNet dataset, 4.9–19.3% on the FMB dataset, and 0.4–7.3% on the ZJU dataset, compared to other SoTA methods. These results underscore the versatility of our network in handling diverse data types. It is important to note that since the "bicycle" category is not included in the test set of the FMB dataset, and we report the mIoU metrics excluding the "bicycle" category.

Qualitative comparisons on the KITTI Road, Cityscapes, and MFNet datasets are presented in Figs. 3–5. The dual-branch feature fusion design of RoadFormer+ enables effective capture of both local and global contexts, thereby outperforming previous single-branch heterogeneous feature fusion approaches. Our

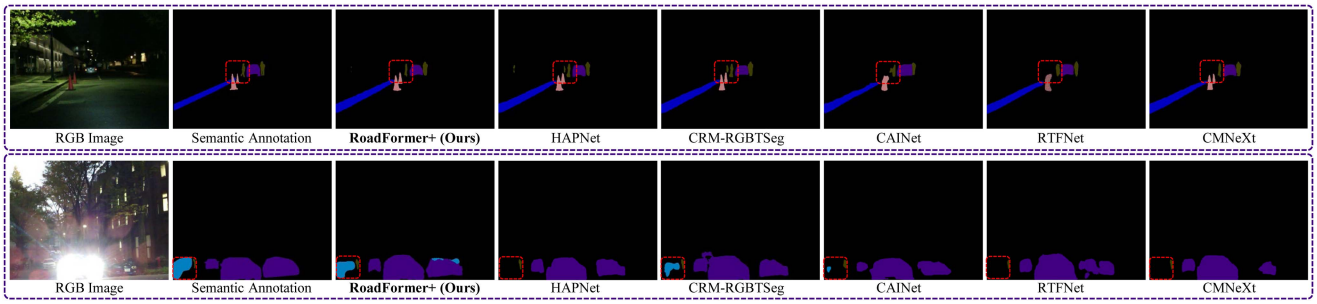

Fig. 5. Qualitative comparisons between our proposed RoadFormer+ and other SoTA networks on the MFNet test set, with significantly improved regions highlighted in red dashed boxes.

TABLE IX
ABLATION STUDY ON THE BACKBONE TRAINING STRATEGY WHEN USING CONVNEXT AS THE BACKBONE

Strategy	SYN-UDTIRI IoU (%)↑	MFNet mIoU (%)↑	ZJU mIoU (%)↑	#Params (M)* ↓
Weight-Separating	**92.88**	58.88	**92.54**	206.8
Weight-Sharing	92.87	**58.96**	92.47	**113.7**

*The resolution of the input image is set to 640×352 pixels.

TABLE X
ABLATION STUDY ON THE BACKBONE SELECTION AND THE EFFECTIVENESS OF OUR PROPOSED HFDE

Backbone	GFE	LFE	IoU (%) ↑	#Params (M) ↓
ConvNeXt-B	✓	×	93.05	**123.8**
ConvNeXt-B	×	✓	93.13	124.9
ConvNeXt-B	✓	✓	**93.44**	134.9
DiNAT-B	✓	✓	93.19	136.1
UniRepLKNet-B	✓	✓	93.36	145.5

method not only demonstrates robust performance in comprehensive scene understanding but also excels in delineating detailed boundaries. Additionally, RoadFormer+ exhibits superior capabilities in handling challenging conditions such as darkness and fog, demonstrating its versatility across diverse scenarios. Furthermore, RoadFormer+ consistently delivers robust performance across various illumination conditions. As illustrated in the second row of Fig. 5, RoadFormer+ outperforms all existing data-fusion methods in handling overexposed scenes within the MFNet dataset.

D. Ablation Studies

We conduct ablation studies on the SYN-UDTIRI, MFNet, and ZJU datasets. Our baseline is built upon RoadFormer [1], and all implementation details are consistent with those described in Section IV-B.

1) Effectiveness of HFDE: Building on our previous findings stated in [1] that demonstrated the effectiveness of ConvNeXt [36] in urban scene parsing, we continue to employ it as the backbone in this study. We investigate two backbone training strategies: weight-sharing and weight-separating. The results, presented in Table IX, show that the weight-sharing strategy not only achieves performance comparable to the weight-separating strategy across three RGB-X datasets but also significantly

TABLE XI
ABLATION STUDY ON THE EFFECTIVENESS OF OUR PROPOSED MHFF BLOCK

Feature Fusion Method	SYN-UDTIRI IoU (%)↑	MFNet mIoU (%)↑	ZJU mIoU (%)↑
HFFM + FFRM	93.44	59.34	92.72
HFFM + LFFM + FFRM	93.67	60.13	92.70
GFRM + LFFM + FFRM	93.82	60.51	92.85
GFRM + LFFM + FEIM	**93.91**	**60.91**	**92.89**
GFRM + LFFM* + FEIM	93.45	60.69	92.64
GFRM + LFFM + FEIM$^{\diamond}$	93.76	59.42	92.55

* The feature channel number of LFFM is doubled due to direct duplication.
$^{\diamond}$ $\boldsymbol{Z}_i^h$ and $\boldsymbol{Z}_i^w$ in the FEIM are processed separately without interaction.

reduces the model's parameters by nearly half. This observation calls into question the utility of traditional duplex encoder designs in these applications.

We further validate the effectiveness of our proposed GFE and LFE on the SYN-UDTIRI dataset in terms of heterogeneous feature enhancement. It is evident that using either GFEs or LFEs independently can effectively enhance our model's performance, and their combined use results in an IoU increase of 0.57%. Additionally, we compare ConvNeXt with recently proposed models, including the Transformer-based DiNAT [39] and UniRepLKNet [35], which both employ large-kernel convolutions. The results affirm that ConvNeXt continues to exhibit superior performance compared to other backbones.

2) Effectiveness of the MHFF Block: As illustrated in Table XI, we utilize RoadFormer as the baseline and alternately replace its feature fusion module with components from our proposed MHFF block to validate the efficacy of the dual-branch feature fusion design. First, we maintain RoadFormer's HFFM and FFRM to fuse global and local features, with the results depicted in the first row. As indicated in the second row, we maintain the use of the HFFM for global feature fusion while integrating the proposed LFFM for local feature fusion, resulting in performance improvements on the SYN-UDTIRI and MFNet datasets, while maintaining stability on the ZJU dataset. Subsequently, HFFM is replaced with our proposed GFRM, with results shown in the third row. Finally, FFRM is replaced with the proposed FEIM, with results presented in the fourth row. The experimental results underscore the individual effectiveness and compatibility of our proposed GFRM, LFFM, and FEIM. When fully integrated, these modules significantly enhance RoadFormer+'s performance in processing three types of RGB-X data compared to the original RoadFormer's feature fusion method. The feature fusion method presented in row

four is our proposed MHFF block. To further validate the effectiveness of the channel expansion design in LFFM and the collaborative processing of $\boldsymbol{Z}_i^h$ and $\boldsymbol{Z}_i^w$ in FEIM, additional experiments are conducted. Removing these operations leads to a decline in the overall performance, as demonstrated in rows five and six.

V. Conclusion

This article reviewed designs for heterogeneous feature extraction and fusion strategies and introduced RoadFormer+, a highly efficient, robust, and applicable urban scene parsing network. Breaking down our contributions further, our work contains five key technical advancements: two modules for feature decoupling in the encoding stage, and three new components within the feature fusion module. The effectiveness of each contribution was validated through extensive experiments. RoadFormer+ outperforms other SoTA algorithms across multiple RGB-X scene parsing datasets. Our future work will primarily focus on investigating lightweight algorithms to enhance adaptability to terminal devices.

References

[1] J. Li, Y. Zhang, P. Yun, G. Zhou, Q. Chen, and R. Fan, "RoadFormer: Duplex transformer for RGB-normal semantic road scene parsing," *IEEE Trans. Intell. Veh.*, vol. 9, no. 7, pp. 5163–5172, Jul. 2024.

[2] R. Fan and M. Liu, "Road damage detection based on unsupervised disparity map segmentation," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 11, pp. 4906–4911, Nov. 2020.

[3] C. Min et al., "ORFD: A dataset and benchmark for off-road freespace detection," in *Proc. 2022 IEEE Int. Conf. Robot. Automat. (ICRA)*, 2022, pp. 2532–2538.

[4] R. Fan, U. Ozgunalp, B. Hosking, M. Liu, and I. Pitas, "Pothole detection based on disparity transformation and road surface modeling," *IEEE Trans. Image Process.*, vol. 29, pp. 897–908, 2020.

[5] C. Hazirbas et al., "FuseNet: Incorporating depth into semantic segmentation via fusion-based CNN architecture," in *Proc. 13th Asian Conf. Comput. Vis. (ACCV)*, Springer, 2017, pp. 213–228.

[6] N. Huang et al., "Discriminative unimodal feature selection and fusion for RGB-D salient object detection," *Pattern Recognit.*, vol. 122, 2022, Art. no. 108359.

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

[8] Q. Ha, K. Watanabe, T. Karasawa, Y. Ushiku, and T. Harada, "MFNet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes," in *Proc. 2017 IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, 2017, pp. 5108–5115.

[9] K. Xiang et al., "Polarization-driven semantic segmentation via efficient attention-bridged fusion," *Opt. Exp.*, vol. 29, no. 4, pp. 4802–4820, 2021.

[10] Z. Wu, Y. Feng, C.-W. Liu, F. Yu, Q. Chen, and R. Fan, "S^3M-Net: Joint learning of semantic segmentation and stereo matching for autonomous driving," *IEEE Trans. Intell. Veh.*, vol. 9, no. 2, pp. 3940–3951, Feb. 2024.

[11] Z. Feng, Y. Guo, and Y. Sun, "Segmentation of road negative obstacles based on dual semantic-feature complementary fusion for autonomous driving," *IEEE Trans. Intell. Veh.*, vol. 9, no. 4, pp. 4687–4697, Apr. 2024, doi: 10.1109/TIV.2024.3376534.

[12] J. Zhang, H. Liu, K. Yang, X. Hu, R. Liu, and R. Stiefelhagen, "CMX: Cross-modal fusion for RGB-X semantic segmentation with transformers," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 12, pp. 14679–14694, Dec. 2023.

[13] Z. Zhao et al., "CDDFuse: Correlation-driven dual-branch feature decomposition for multi-modality image fusion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp. 5906–5916.

[14] Y. Lv, Z. Liu, and G. Li, "Context-aware interaction network for RGB-T semantic segmentation," *IEEE Trans. Multimedia*, vol. 26, pp. 6348–6360, 2024.

[15] X. Li, Y. Li, H. Chen, Y. Peng, and P. Pan, "CCAFusion: Cross-modal coordinate attention network for infrared and visible image fusion," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 2, pp. 866–881, Feb. 2024.

[16] S.-J. Park, S. Lee, and K.-S. Hong, "RDFNet: RGB-D multi-level residual feature fusion for indoor semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2017, pp. 4980–4989.

[17] A. Alhaija et al., "Augmented reality meets computer vision: Efficient data generation for urban driving scenes," *Int. J. Comput. Vis.*, vol. 126, pp. 961–972, 2018.

[18] M. Cordts et al., "The cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 3213–3223.

[19] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2015, pp. 3431–3440.

[20] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 2881–2890.

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

[22] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 4510–4520.

[23] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.

[24] R. Strudel, R. Garcia, I. Laptev, and C. Schmid, "Segmenter: Transformer for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2021, pp. 7262–7272.

[25] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2017, pp. 6000–6010.

[26] S. Zheng et al., "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021, pp. 6881–6890.

[27] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Representations (ICLR)*, 2020.

[28] B. Cheng et al., "Per-pixel classification is not all you need for semantic segmentation," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2021, pp. 17864–17875.

[29] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Masked-attention mask transformer for universal image segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 1290–1299.

[30] Y. Sun, W. Zuo, and M. Liu, "RTFNet: RGB-thermal fusion network for semantic segmentation of urban scenes," *IEEE Robot. Automat. Lett.*, vol. 4, no. 3, pp. 2576–2583, Jul. 2019.

[31] Y. Feng et al., "SNE-RoadSegV2: Advancing heterogeneous feature fusion and fallibility awareness for freespace detection," 2024, *arXiv:2402.18918*.

[32] Y. Yang, C. Shan, F. Zhao, W. Liang, and J. Han, "On exploring shape and semantic enhancements for RGB-X semantic segmentation," *IEEE Trans. Intell. Veh.*, vol. 9, no. 1, pp. 2223–2235, Jan. 2024.

[33] H. Wang et al., "RGB-depth fusion GAN for indoor depth completion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 6209–6218.

[34] W. Zhou, S. Dong, M. Fang, and L. Yu, "CACFNet: Cross-modal attention cascaded fusion network for RGB-T urban scene parsing," *IEEE Trans. Intell. Veh.*, vol. 9, no. 1, pp. 1919–1929, Jan. 2024.

[35] X. Ding et al., "UniRepLKNet: A universal perception large-kernel ConvNet for audio video point cloud time-series and image recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2024, pp. 5513–5524.

[36] Z. Liu, H. Mao, C. -Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 11976–11986.

[37] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 770–778.

[38] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2021, pp. 10012–10022.

[39] A. Hassani and H. Shi, "Dilated neighborhood attention transformer," 2022, *arXiv:2209.15001*.
[40] Z. Wuand et al., "Lite transformer with long-short range attention," in *Proc. Int. Conf. Learn. Representations (ICLR)*, 2020.
[41] X. Chen et al., "Bi-directional cross-modality feature propagation with separation-and-aggregation gate for RGB-D semantic segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Springer, 2020, pp. 561–577.
[42] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 7132–7141.
[43] J. Fu et al., "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 3146–3154.
[44] I. O. Tolstikhin et al., "MLP-Mixer: An all-MLP architecture for vision," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2021, pp. 24261–24272.
[45] Q. Hou, D. Zhou, and J. Feng, "Coordinate attention for efficient mobile network design," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021, pp. 13713–13722.
[46] S. Woo et al., "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 3–19.
[47] A. Dosovitskiy et al., "CARLA: An open urban driving simulator," in *Proc. Conf. Robot Learn. (CoRL)*, PMLR, 2017, pp. 1–16.
[48] A. Geiger et al., "Vision meets robotics: The KITTI dataset," *Int. J. Robot. Res.*, vol. 32, no. 11, pp. 1231–1237, 2013.
[49] L. Lipson, Z. Teed, and J. Deng, "RAFT-Stereo: Multilevel recurrent field transforms for stereo matching," in *Proc. 2021 IEEE Int. Conf. 3D Vis. (3DV)*, 2021, pp. 218–227.
[50] M. Menze and A. Geiger, "Object scene flow for autonomous vehicles," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2015, pp. 3061–3070.
[51] C.-W. Liu et al., "Playing to vision foundation model's strengths in stereo matching," *IEEE Trans. Intell. Veh.*, 2024.
[52] J. Liu et al., "Multi-interactive feature learning and a full-time multi-modality benchmark for image fusion and segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2023, pp. 8115–8124.
[53] Y. Bowen et al., "DFormer: Rethinking RGBD representation learning for semantic segmentation," in *Proc. Int. Conf. Learn. Representations (ICLR)*, 2024.
[54] Y. Feng, B. Xue, M. Liu, Q. Chen, and R. Fan, "D2NT: A high-performing depth-to-normal translator," in *Proc. 2023 IEEE Int. Conf. Robot. Automat. (ICRA)*, 2023, pp. 12360–12366.
[55] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. Int. Conf. Learn. Representations (ICLR)*, 2019.
[56] E. Xie et al., "SegFormer: Simple and efficient design for semantic segmentation with transformers," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 12077–12090.
[57] Y. Zhu et al., "Improving semantic segmentation via video propagation and label relaxation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 8856–8865.
[58] J. Zhang et al., "Delivering arbitrary-modal semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp. 1136–1147.
[59] U. Shin et al., "Complementary random masking for RGB-thermal semantic segmentation," 2023, *arXiv:2303.17386*.
[60] J. Li et al., "HAPNet: Toward superior RGB-thermal scene parsing via hybrid, asymmetric, and progressive heterogeneous feature fusion," *arXiv:2404.03527*, 2024.

Jianxin Huang is currently a Research Assistant with Tongji University, Shanghai, China. His research interests include computer vision and deep learning.

Jiahang Li is currently working toward the M.Sc. degree with Tongji University, Shanghai, China. His research interests include computer vision and deep learning.

Ning Jia is currently an Assistant Professor with Tongji University, Shanghai, China. His research interests include computer vision and pattern recognition.

Yuxiang Sun (Member, IEEE) is currently an Assistant Professor with the Department of Mechanical Engineering, City University of Hong Kong, Hong Kong. His research interests include Robotics and AI.

Chengju Liu is currently a Professor with Tongji University, Shanghai, China. Her research interests include motion control of legged robots, embodied AI, and vision-and-language navigation.

Qijun Chen (Senior Member, IEEE) is currently a Professor with Tongji University, Shanghai, China. His research interests include computer vision, deep learning, and robotics.

Rui Fan (Senior Member, IEEE) is currently a Professor with Tongji University, Shanghai, China. His research interests include computer vision, deep learning, and robotics.