MMFSeg: Multi-Structure Multi-Feature Fusion for Segmentation of Road Potholes

Zhen Feng, Yanning Guo, Rui Fan, and Yuxiang Sun

***Abstract*—Road pothole segmentation is important for the driving safety of autonomous vehicles, especially in unstructured or rural environments. Recently, many effective multi-modal fusion networks have been proposed for road pothole segmentation. However, most of them adopt two encoders with the same type of structure, such as only convolutional neural network (CNN) or only Transformer, to extract features from different modalities. This overlooks the fact that the information richness of features extracted from different modalities varies across the types of features, such as CNN features or self-attention features. To provide a solution to this issue, we design a novel RGB-Disparity segmentation network, named MMFSeg, by adopting the two types of structures as encoders. Specifically, we adopt CNN and Transformer as encoders to extract two types of features from each modality, and also propose a late-fusion multi-feature alignment fusion module to fuse the two types of features with different numbers of channels. Experimental results demonstrate that our network outperforms well-known networks, and can trade-off between accuracy and efficiency.**

***Note to Practitioners*—This study is driven by the challenge of segmenting road potholes to ensure the safety of autonomous driving. Our proposed network can utilize different structures to extract diverse feature information from each modality of data, thereby achieving more accurate road pothole segmentation results. Our proposed fusion module effectively fuses diverse features from each modality through a late fusion strategy. Our method validates effectiveness on multiple datasets. This work contributes to the safety of autonomous vehicles by enhancing their driving performance under poor road conditions. Our network achieves superior performance compared with existing networks, thereby enabling more effective deployment of downstream tasks, such as path planning and navigation, in autonomous driving systems.**

***Index Terms*—Pothole segmentation, multi-modal fusion, autonomous vehicles, convolution-transformer structure.**

I. Introduction

ROAD potholes, usually appearing on roads without maintenance [1], pose a great threat to the driving safety of vehicles [2]. Road potholes not only cause bumps, but also lead to accidents if vehicle speed is fast and potholes are large. So, accurate detection of potholes is essential to autonomous driving [3]. Semantic segmentation has been extensively used to detect objects at the pixel level [4], [5], [6]. So, detection of road potholes based on semantic segmentation has received great attention [7].

Recently, many single-modal networks have demonstrated effective road pothole segmentation results [8], [9]. However, due to the limitation of single-modal data to environmental changes, the performance of these networks degrades as the data quality degrades. For example, poor lighting conditions at nighttime may decrease useful information in RGB images, hence hindering the segmentation accuracy [10]. So, single-modal networks could not well work in complex and changing road environments. To address this issue, multi-modal fusion networks have been applied to this area [11], [12]. Considering depth differences between potholes and road surfaces, some methods have been proposed by fusing RGB and depth data (e.g., point clouds [13] and depth images [14]).

Although these multi-modal fusion networks present acceptable results, they employ two encoders with the same structure to extract features of the same type from different modalities at the same stage [15], [16], ignoring the effects between different modalities and different types of features. For example, using convolutional neural network (CNN)-based encoders to extract convolutional features from RGB images and disparity images [12], or using Transformer-based encoders to extract self-attention features from RGB images and depth images [14]. We argue that extracting the same type of features, such as convolutional features or self-attention features, from different modalities may not fully utilize the information from each modality.

To provide a solution to the above issues, we design a novel multi-modal fusion network, which adopts a CNN-based encoder and a Transformer-based encoder to extract two different types of features from each modality. To fuse different types of features, we design a Multi-feature Alignment Fusion (MAF) module with the late fusion strategy to fuse two features that have different numbers of channels. The experimental results show that our proposed network can better fuse two different types of features. Moreover, the superiority of our network is demonstrated by comparing with existing well-known multi-modal fusion networks. Our code is open-

Received 2 January 2025; revised 8 May 2025 and 27 June 2025; accepted 17 September 2025. Date of publication 23 September 2025; date of current version 30 October 2025. This article was recommended for publication by Associate Editor W. Zhang and Editor D. Song upon evaluation of the reviewers' comments. This work was supported by the City University of Hong Kong under Grant 9610675. *(Corresponding author: Yuxiang Sun.)*

Zhen Feng and Yuxiang Sun are with the Department of Mechanical Engineering, City University of Hong Kong, Hong Kong (e-mail: zhen.feng@cityu.edu.hk; yx.sun@cityu.edu.hk).

Yanning Guo is with the Department of Control Science and Engineering, Harbin Institute of Technology, Harbin 150001, China (e-mail: guoyn@hit.edu.cn).

Rui Fan is with the College of Electronic and Information Engineering, Tongji University, Shanghai 201804, China (e-mail: rui.fan@ieee.org).

Digital Object Identifier 10.1109/TASE.2025.3613629

1558-3783 © 2025 IEEE. All rights reserved, including rights for text and data mining, and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: CITY UNIV OF HONG KONG. Downloaded on November 05,2025 at 04:15:11 UTC from IEEE Xplore. Restrictions apply.

TABLE I
CATEGORIZATION OF THE SINGLE-MODAL SEGMENTATION NETWORKS BY THEIR BACKBONES AND APPLICATION DOMAINS

Network	Backbone	Application domain
UNet [19]	CNN	Medical
Deeplab [20]	ResNet	General
Deeplabv3+ [21]	ResNet	General
InspectionNet++ [24]	ResNet	Crack
ConvUNeXt [22]	ConvNeXt	Medical
Meta-UNet [27]	Transformer	Medical
SegFormer [29]	Transformer	General
SeaFormer [30]	Transformer	General
SwinPA-Net [31]	Swin-Transformer	Medical
TransAttUnet [32]	Transformer	Medical
MorFormer [33]	Transformer	Crack

sourced.[1] The contributions of this paper are summarized as follows:

1) We design a novel RGB-Disparity (RGB-D) network for pothole segmentation, named MMFSeg, which adopts CNN and Transformer as encoders to extract two types of features from each modality.
2) We design the MAF module to fuse the two types of features that have different numbers of channels.
3) We conduct comparative experiments to compare our MMFSeg with well-known networks. The results demonstrate the superiority of our network.

This paper is structured as follows. Section II reviews the related work. Section III describes the details of our proposed method. Section IV discusses the experimental results. The last section concludes this work and discusses the future work.

II. RELATED WORK

A. Single-Modal Segmentation Networks

CNN and Transformer are commonly used as backbones to design semantic segmentation networks [17], [18]. Ronneberger et al. [19] proposed the U-shaped semantic segmentation network, UNet, with the encoder-decoder architecture. Chen et al. proposed Deeplab [20] and Deeplabv3+ [21] with the atrous convolution in the encoder-decoder architecture. Han et al. [22] designed ConvUNeXt by combining UNet and ConvNeXt [23] for medical image segmentation. Feng et al. [24] adopted ResNet [25] as an encoder to design InspectionNet++ for detecting cracks on concrete infrastructure. Dosovitskiy et al. [26] introduced the Transformer from natural language processing, which is then widely used as the encoder of semantic segmentation networks. Wu et al. [27] designed a multi-Scale efficient transformer attention mechanism with a U-shaped network for the segmentation of polyp segmentation. Recently, Many segmentation networks based on Transformer have been proposed [28], such as SegFormer [29] and SeaFormer [30]. For medical image segmentation, Du et al. [31] designed SwinPA-Net with Swin-Transformer as the encoder, Chen et al. [32] designed TransAttUnet by combining Transformer and UNet. Guo et al. [33] used the Transformer as encoders and introduced a morphology-aware mechanism to design MorFormer for the pavement crack segmentation. The above networks are categorized in Tab. I by their backbones and application domains.

B. Multi-Modal Fusion Segmentation Networks

Element-wise addition and concatenation are commonly used to fuse feature maps extracted from different modalities of data. However, Feng et al. [11] found that simple element-wise addition and concatenation sometimes lead to performance degradation when fusing features from two modalities with inconsistent information. Element-wise multiplication is often used to capture the interactions between modalities. However, it may result in information loss in modeling high-dimensional dependencies [34]. Element-wise maximum is also another commonly used fusion method. However, it is sensitive to noise and may lead to information loss [35], [36]. So, these methods may not be appropriate to be used to fuse features with obvious noises. Average value fusion is a weighted fusion method. It ignores the properties of each modality [36]. Sun et al. [37] designed RTFNet that adopts element-wise addition to fuse the feature maps from RGB and thermal images. Fan et al. [38] designed RoadSeg that adopts element-wise addition to fuse the feature maps from RGB and depth images. To improve fusion performance, some researchers resort to using attention mechanisms. Zhou et al. [39] designed a cross-modality awareness module in FRNet to fuse the features from different modalities. Liang et al. [40] designed an explicit attention-enhanced fusion module based on attention mechanisms to fuse RGB-Thermal (RGB-T) features. Cai et al. [41] designed DHFNet to fuse RGB-T features with an adaptive attention-filtering fusion module. Zhang et al. [42] designed a multi-modal fusion knowledge distillation framework based on the channel attention mechanism.

In addition, some researchers resort to using specially-designed fusion strategies. Zhou et al. [43] proposed GMNet for RGB-T segmentation with two fusion modules respectively for shallow features and deep features. Zhou et al. [44] designed a multiple-strategy fusion module to fuse different features by different strategy, such as element-wise addition, element-wise multiplication, maximum pixel value, and average pixel values. Feng et al. [45] proposed an adaptive-mask module to avoid the effect of invalid information in depth images, and introduced adaptive weights to fuse the features of RGB-Depth (RGB-D) images. Zhou et al. [46] designed a demand-modal adaptive module to adaptively determine the ratio of the integrated cross-modal information for RGB-D feature fusion. Feng et al. [47] designed a three-stage fusion module with spatial attention and channel attention to fuse features of RGB-D images. In the staged fusion module, a holistic attention module is designed to discriminate inherent

[1]https://github.com/lab-sun/MMFSeg

Authorized licensed use limited to: CITY UNIV OF HONG KONG. Downloaded on November 05,2025 at 04:15:11 UTC from IEEE Xplore. Restrictions apply.

differences between different features from different modalities. A heterogeneous feature contrast descriptor is designed to capture shared and distinct characteristics of different features. In the last stage, the features are adaptively fused by the adaptive weights. The staged fusion module enhances the comprehensiveness of feature fusion for freespace detection. Zhou et al. [48] designed an RGB-T fusion network MMSM-CNet with the modal memory sharing (MMS) module that adopts the staged fusion technique. They separately extract the contour and skeleton of a target, and finally obtains the complementary information through the morphological complementary modules, which optimizes targets of different morphologies at various scales. Zhou et al. [49] adopted a method that emphasizes feature contrast and difference inversion for multi-modal feature fusion. And then, they refined high-level semantic information and complemented the modality with clustered instance regularization. Feng et al. [11] proposed a residual-guided fusion module placed at the decoder stage to complement the missing features of RGB images for obstacles segmentation. Li et al. [50] designed a heterogeneous feature synergy block to fuse features of two modalities with a feature fusion stage and a feature recalibration stage. Huang et al. [51] divided the features of each modality into global features and local features, and then adopted a global feature recalibration module and a local feature fusion module to fuse the features from two modalities.

C. Pothole Segmentation Networks

Many pothole segmentation networks based on single modal of data have been proposed. RGB images are commonly used in pothole segmentation tasks. Han et al. [52] designed a reflection attention unit (RAU) and introduced the unit into a fully convolutional network to segment water hazards using only RGB images. The proposed RAU is designed to capture the light-reflective properties of water surfaces in water hazards across different image regions based on a reflection correspondence mechanism. RAU captures the difference between water hazards and roads, which improves the accuracy of water hazards segmentation. Huang et al. [53] proposed a plug-and-play embranchment aggregation and detail enhancement module to improve the performance of pothole segmentation and detection. Zhou et al. [54] designed LightCrackNet for crack segmentation based on split exchange convolution that decomposes the features into high- and low-resolution features, and utilizes the pooling layer to reduce the size of the low-resolution features to decrease the data volume. The convolution makes full use of information of different scales and significantly reduces the number of parameters and computational complexity, enabling realization of high efficiency and lightweight. Dong et al. [9] proposed CRAM-Seg-CapsNet with neural capsules that extract vector features for pothole segmentation. Since cracks are similar to potholes, some researchers also studied crack segmentation. Liu et al. [55] introduced atrous spatial pyramid pooling and coordinate attention blocks into UNet to design a network for pothole and crack segmentation. Xu et al. [56] proposed a wall crack segmentation network with a model-agnostic meta-learning method that was deployed in a drone. Ma et al. [57] proposed an unsupervised network UP-CrackNet for the detection of road cracks. UP-CrackNet adopts undamaged road images with random masks to train an image generator to detect the cracks in damaged road images.

The ability of thermal images to capture infrared radiation from objects enables their application in pothole segmentation tasks. Aparna et al. [10] built a dataset with thermal images for pothole segmentation. They used thermal cameras to capture images of potholes on sunny days and at night, respectively. The authors also designed a convolutional network for this task.

Many multi-modal fusion methods have been proposed to improve the performance of pothole segmentation. RGB-D images are commonly used in pothole segmentation tasks. Fan et al. [12] designed RGB-disparity fusion networks, AAUNet and AARTFNet, by placing channel attention modules (CAMs) and dual attention modules (DAMs) between the encoder and decoder stages of UNet and RTFNet. The authors released an RGB-disparity dataset with 600 pairs of images for pothole segmentation. They adopted the CAMs to select some more important channels that are more crucial for segmenting potholes by adjusting different weights for different channels. The DAMs are placed at the last two stages of the encoders due to the limitation of the memory consumed. The DAMs are used to learn the weight distribution of the feature maps in both spatial and channel dimensions, selecting the important features for better results. Feng et al. [2] designed MAFNet by combining Transformer [26] and ResNet [25] as the encoder for the segmentation of road potholes. They used ResNet to extract local features from low-level feature maps, and used Transformer to extract global features from the last high-level feature maps. The combination of the local and global features improves characterization learning for multi-modal road pothole segmentation. The authors also designed two fusion modules based on channel attention and dual attention to fuse the features of RGB images and disparity images. Feng et al. [58] proposed a multi-modal fusion knowledge distillation framework for road-pothole segmentation. They designed a channel and position-wise distillation module to transfer knowledge. Feng et al. [14] designed PotCrackSeg for the segmentation of potholes and cracks, which adopts a proposed dual semantic-feature complementary fusion module to fuse the features from RGB and depth images.

The paradigm of fusing RGB images with LiDAR point clouds is also applied to road pothole segmentation. Li et al. [13] designed a fusion module based on the attention mechanism to fuse features of RGB images and point clouds for pothole segmentation. They designed residual mapping structure and point attention to fuse the features of RGB images and point clouds.

D. Difference From Existing Work

Although the existing multi-modal networks for pothole segmentation have achieved effective results, most of them adopt two encoders with the same structure (i.e., only CNN

Authorized licensed use limited to: CITY UNIV OF HONG KONG. Downloaded on November 05,2025 at 04:15:11 UTC from IEEE Xplore. Restrictions apply.

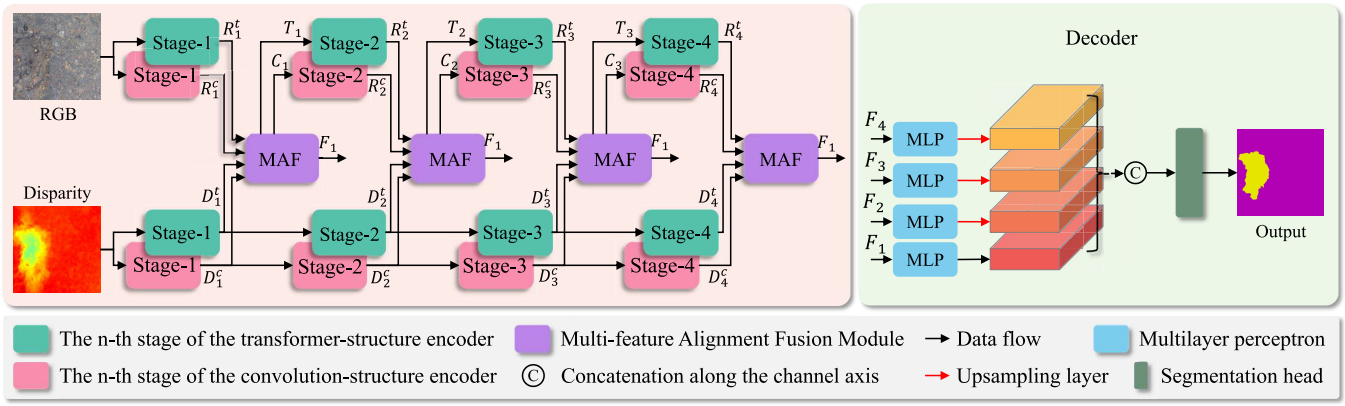


Fig. 1. The overall architecture of our proposed MMFSeg. It adopts the encoder-decoder architecture: two convolution-based encoders, two Transformer-based encoders, and one decoder. We adopt ConvNeXt-T as convolutional encoders and SegFormer-b0 as Transformer encoders.

or only Transformer) to extract features. The main difference between ours and the existing works is that we combine both CNN and Transformer for feature extraction. This could fully take the advantages of CNN and Transformer.

III. The Proposed Network

A. The Overall Architecture

Our MMFSeg adopts CNN and Transformer to extract features from two modalities of data at the same time. Since the features extracted by the two structures are with different types and usually have different numbers of channels, we design the MAF module to fuse the two types of features. Our MMFSeg adopts the encoder-decoder structure that contains two CNN-based encoders (denoted as convolutional encoders), two Transformer-based encoders (denoted as transformer encoders), and one decoder. The MAF modules are placed to fuse the feature maps from each encoder stage and the outputs are fed into the decoder. The overall structure of our network is shown in Fig. 1.

We adopt ConvNeXt-T [23] as the convolutional encoders and SegFormer-B0 [29] as the transformer encoders. In the encoder stage, there is an RGB stream, a disparity stream, and four MAF modules. The RGB stream contains a convolutional encoder and a transformer encoder, which is used to extract features from RGB images. The disparity stream has a similar structure as the RGB stream.

Each encoder has four stages for feature extraction. The output of the n-th stage of the RGB/disparity convolutional encoder is denoted as R_n^c/D_n^c, and the n-th stage of the RGB/disparity transformer encoder is denoted as R_n^t/D_n^t, where $n \in [0, 4]$. The outputs of each stage of the disparity stream are fed into the MAF module and the following stages. But the outputs of each stage of the RGB stream are only fed into the MAF module. In the RGB stream, except for the first stages of the encoders, the inputs of the $(n+1)$-th stages of the encoders are the outputs of the n-th MAF module, where $n \in [1, 3]$. The n-th MAF module has 3 outputs: transformer features T_n, convolutional features C_n, and fusion features F_n, where $n \in [1, 4]$. The transformer features T_n are fed into the RGB transformer encoder, and the convolutional features C_n are fed into the RGB convolutional encoder. Note that the last MAF module only outputs fused features. All of the fused features from all the MAF modules are fed into the decoder. We detail the pipeline of the encoder in a pseudocode, which is shown in Algorithm 1.

Algorithm 1 The Encoder

Data: RGB images I_R, disparity imgaes I_D
Result: Fusion features: F_1, F_2, F_3, F_4

$R_{in}^t \leftarrow I_R$, $R_{in}^c \leftarrow I_R$; // RGB stream inputs
$D_{in}^t \leftarrow I_D$, $D_{in}^c \leftarrow I_D$; // Disparity stream inputs
for $i \leftarrow 1$ **to** 4 **do** // Stages
  $R_i^c \leftarrow C_R^i(R_{in}^c)$; // RGB convolutional encoder
  $R_i^t \leftarrow T_R^i(R_{in}^t)$; // RGB transfromer encoder
  $D_i^c \leftarrow C_D^i(D_{in}^c)$; // Disparity convolutional encoder
  $D_i^t \leftarrow T_D^i(D_{in}^t)$; // Disparity transfromer encoder
  $T_i, C_i, F_i \leftarrow \text{MAF}(R_i^c, R_i^t, D_i^c, D_i^t)$; // MAF
  $R_{in}^t \leftarrow T_i$, $R_{in}^c \leftarrow C_i$;
  $D_{in}^t \leftarrow D_i^t$, $D_{in}^c \leftarrow D_i^c$;

B. The MAF Module

Due to the structural differences in convolutional encoders and transformer encoders, the features extracted by the same-level stages have different numbers of channels and different types. The numbers of channels of the features extracted by the convolutional encoder are 96, 192, 384, and 768, respectively. The numbers of channels of the features extracted by the transformer encoder are 32, 64, 160, and 256, respectively. To address the issue of fusing features containing different numbers of channels and different types, we design the MAF module in our MMFSeg. The structure of the MAF module is shown in Fig. 2. Algorithm 2 presents the details of the operations in the MAF module in pseudo-code.

The MAF module contains three blocks: spatial block, alignment block, and channel block. The spatial block is designed to modify the spatial weights of two different types

Authorized licensed use limited to: CITY UNIV OF HONG KONG. Downloaded on November 05,2025 at 04:15:11 UTC from IEEE Xplore. Restrictions apply.

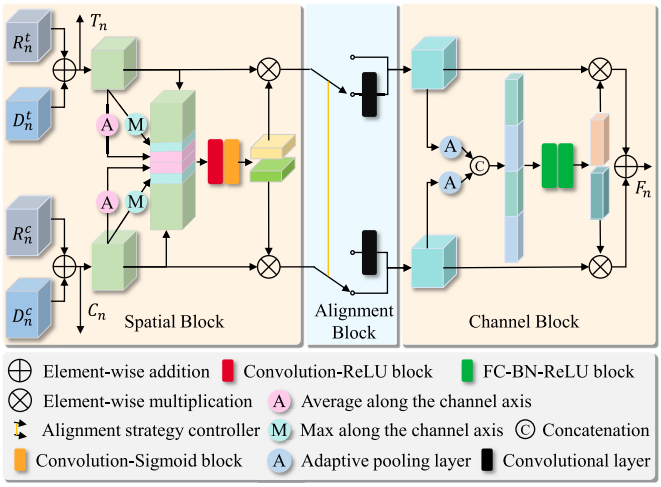


Fig. 2. The structure of our proposed MAF module. FC and BN refer to the fully connected layer and batch normalization layer.

Algorithm 2 The n-Th MAF Module

Data: RGB CNN features R_n^c, RGB Transformer features R_n^t, disparity CNN features D_n^c, disparity Transformer features D_n^t
Result: CNN features C_n, Transformer features T_n, Fusion features F_n

// **The following is the spatial block**
$C_n \leftarrow R_n^c + D_n^c$, $T_n \leftarrow R_n^t + D_n^t$;
$A_n^c \leftarrow ave(C_n)$, $M_n^c \leftarrow max(C_n)$;
$A_n^t \leftarrow ave(T_n)$, $M_n^c \leftarrow max(T_n)$;
$S_n \leftarrow Conc(T_n, A_n^t, M_n^t, A_n^c, M_n^c, C_n)$;
$[w_s^t; w_s^c] \leftarrow ConvS\big(ConvR(S_n)\big)$;
$T_n^s \leftarrow T_n \otimes w_s^t$, $C_n^s \leftarrow C_n \otimes w_s^c$;
// **The following is the alignment block**
if *C2T* **then**
  $T_n^a \leftarrow T_n^s$;
  $C_n^a \leftarrow Conv(C_n^s)$;
else
  $T_n^a \leftarrow Conv(T_n^s)$;
  $C_n^a \leftarrow C_n^s$;
// **The following is the channel block**
$P_n^t \leftarrow adapool(T_n^a)$, $P_n^c \leftarrow adapool(T_n^a)$;
$P \leftarrow Conc(P_n^t, P_n^c, P_n^t, P_n^c)$;
$[w_c^t; w_c^c] \leftarrow fbr\big(fbr(P)\big)$;
$T_n^c \leftarrow T_n^a \otimes w_c^t$, $C_n^c \leftarrow C_n^a \otimes w_c^c$;
$F_n \leftarrow T_n^c + C_n^c$;

of features. In the spatial block, the same type of features are first fused with element-wise addition, that is, $T_n = R_n^t + D_n^t$ and $C_n = R_n^c + D_n^c$, where $n \in [1,4]$. The fusion results T_n and C_n are fed into the next stages of the RGB stream. Secondly, the average and maximum values of T_n and C_n along the channel axis are calculated. The processes of the average values and maximum values are denoted as $ave(\cdot)$ and $max(\cdot)$. The average results (A_n^c and A_n^t), maximum results (M_n^c and M_n^t), and the original T_n and C_n are concatenated along the channel axis. Thirdly, the concatenated result is fed into two consecutive blocks: a convolution-ReLU block denoted as $convR(\cdot)$ containing a convolutional layer and a ReLU layer, as well as a convolution-Sigmoid block denoted as $convS(\cdot)$ containing a convolutional layer and a Sigmoid layer. The convolution-Sigmoid block generates 2-channel spatial weight maps with the same resolution as the concatenated result. Finally, the spatial weight maps are divided to fuse with T_n and C_n through element-wise multiplication, respectively. The fusion results (T_n^s and C_n^c) are fed into the alignment block.

The alignment block, containing an alignment strategy controller and two convolutional layers $conv(\cdot)$, is used to resize the number of channels to be the same for both features. The controller controls the alignment strategies: 1) Align transformer features to convolutional features (T2C); 2) Align convolutional features to transformer features (C2T). In the T2C strategy, the T_n^s is fed into a convolutional layer to resize the number of channels to that of C_n^c. But there are no operations for C_n^c. In the C2T strategy, the C_n^c is fed into a convolutional layer to resize the number of channels to that of T_n^s. But there are no operations for T_n^s. The ablation study results demonstrate that the C2T strategy offers the better trade-off between performance and efficiency, and thus is adopted in our MMFSeg framework. The outputs of the alignment block (T_n^a and C_n^a) are fed into the channel block.

In the channel block, the inputs are first fed into adaptive pooling layers (denoted as $adapool(\cdot)$). Then, the outputs of the adaptive pooling layers are concatenated along the channel axis. We repeat each output twice to provide more information. Thirdly, the concatenated result is fed into two consecutive blocks (denoted as $fbr(\cdot)$) to generate the channel weights for each input of the channel block. Each block contains a fully connected layer, a batch normalization layer, and a ReLU layer. Then, the channel weights and the inputs of the channel block are fused with element-wise multiplication. Finally, the multiplication results are fused with element-wise addition, and the fusion result F_n is fed into the decoder.

C. The Decoder

In the decoder, we adopt 4 Multilayer Perceptron (MLP) to restore information from the inputs (i.e., F_1, F_2, F_3, and F_4). The outputs of all MLPs have the same number of channels. The outputs of the MLP for F_2, F_3, and F_4 are fed into upsampling layers to resize the resolution to that of F_1. Next, the same-resolution feature maps are fed into a segmentation head after concatenating along the channel axis. The segmentation head contains a convolutional layer, a batch normalization layer, and a ReLU layer, as well as a dropout layer and a convolutional layer. The segmentation head finally generates the segmentation maps that have the same resolution as the input RGB-Disparity images.

IV. Experimental Results and Discussions

A. The Dataset

We use the `Pothole-600` dataset [12] in our experiments. The dataset contains RGB images captured by a ZED camera, as well as disparity images generated by the disparity transformation algorithm [59]. There are 600 pairs of images that are

Authorized licensed use limited to: CITY UNIV OF HONG KONG. Downloaded on November 05,2025 at 04:15:11 UTC from IEEE Xplore. Restrictions apply.

divided into three subsets: training set (240 pairs), validation set (180 pairs), and testing set (180 pairs). Each pair of images is provided with a hand-labeled ground truth.

We also use the NO-4K dataset [58] for comparative experiments. The NO-4K dataset is collected from rural environments for pothole segmentation. This dataset contains a large number of images of potholes with puddled water, as well as blurred images caused by fast vehicle speed. The dataset contains 3,745 pairs of RGB and disparity images with the 576×1024 resolution. There are 745 manually-labeled images, which are divided into a validation set with 245 images and a testing set with 500 images. The remaining 3,000 images are formed as the training set with generated labels. The authors used the manually labeled 245 images and 600 images in the whole Pothole-600 to train MAFNet [2]. The well-trained MAFNet is used to generate the pothole mask of the remaining 3,000 images with a confidence threshold of over 0.5. The generated labels contain noises in certain areas (e.g., wet or blurred regions), which is one of the challenges in the NO-4K dataset.

B. Training Details

We implement our MMFSeg with PyTorch 1.12 and train our network with an NVIDIA RTX3090 graphics card (24GB GPU RAM). The parameters of the transformer encoder are initialized with the strategy from [29]. The parameters of the convolutional encoder are initialized with the pre-trained weights provided by [23]. Other parameters are initialized with the PyTorch default method. The AdamW optimizer with 6×10^{-5} of the initial learning rate is used to train our network and its variants. The learning rate is increased from 0 through a 10-epoch warm-up, and then decreased after the 10-th epoch with a polynomial learning rate decay. During training, we apply random clipping and cropping to augment images before feeding them into the network. We treat the unlabeled pixels as *background* class that is also segmented.

C. Evaluation Metrics

We use the metrics, Precision (Pre), Recall (Rec), F-score (F1), and Intersection over Union (IoU), to evaluate the performance of all the networks. They are calculated as follows:

$$\mathrm{Pre}_c = \frac{|\{x|x \in \mathrm{TP}_c\}|}{|\{x|x \in \mathrm{TP}_c \cup \mathrm{FP}_c\}|}, \tag{1}$$

$$\mathrm{Rec}_c = \frac{|\{x|x \in \mathrm{TP}_c\}|}{|\{x|x \in \mathrm{TP}_c \cup \mathrm{FN}_c\}|}, \tag{2}$$

$$\mathrm{F1}_c = \frac{2 \times \mathrm{Pre}_c \times \mathrm{Rec}_c}{\mathrm{Pre}_c + \mathrm{Rec}_c}, \tag{3}$$

$$\mathrm{IoU}_c = \frac{|\{x|x \in \mathrm{TP}_c\}|}{|\{x|x \in \mathrm{TP}_c \cup \mathrm{FP}_c \cup \mathrm{FN}_c\}|}, \tag{4}$$

where c refers to class (i.e., background and potholes), TP_c, FP_c, and FN_c refer to the true-positive area, false-positive area, and false-negative area of class c. x refers to the pixels in the areas. We calculate the four metrics for each class and mean values (i.e., mPre, mRec, mF1, and mIoU) across the two classes (i.e., pothole and backgroud) to evaluate the performance. In addition, we evaluate the efficiency of the networks in terms of the average runtime cost and frame-per-second (FPS) for each image.

TABLE II
THE RESULTS (%) OF SELECTED VARIANTS FROM THE ABLATION STUDY ON ENCODER

Variant	mPre	mAcc	mF1	mIoU	RTX3090	
					ms	FPS
C2T-C0T0	94.12	92.95	93.53	88.31	28.61	34.95
C2T-C0T4	94.21	93.52	93.86	88.85	68.59	14.58
T2C-C1T1	93.68	93.40	93.54	88.32	38.06	26.27
T2C-C2T3	94.17	92.89	93.52	88.30	73.51	13.60

D. Ablation Studies

1) Ablation on Encoder: We conduct ablation studies to find the optimal combination between the transformer encoder and the convolutional encoder. We design variants by using Segformer with different layers, such as Segfomer-B0 (T0), Segfomer-B1 (T1), Segfomer-B2 (T2), Segfomer-B3 (T3), Segfomer-B4 (T4), Segfomer-B5 (T5), and ConvNeXt with different layers, such as ConvNeXt-T (C0), ConvNeXt-S (C1), ConvNeXt-B (C2), ConvNeXt-L (C3). Firstly, we fix the convolutional encoder and change the transformer encoder from T0 to T5 to design variants. Secondly, we fix the transformer encoder and change the convolutional encoder from C0 to C3. Thirdly, we respectively remove the transformer encoder (NT) or convolutional encoder (NC) to design variants. In addition, we also change the alignment strategy (i.e., T2C and C2T) of the MAF module to design variants. We use C2T-C0T0 to refer to the variant that adopts T0 and C0 as encoders and employs the C2T alignment strategy. Note that the C3T4 and C3T5 variants cannot be trained on the NVIDIA RTX3090 due to the limited GPU memories.

The results of the four variants are shown in Fig. 3. We can see that the distribution of the mF1 and mIoU results is similar within C2T and T2C. The C2T/T2C variants refer to the variants adopting the C2T/T2C alignment strategy. The results demonstrate that the appropriate combination of different encoders can yield superior results. In both C2T and T2C, the dual-structured variants achieve better results than the single-structured variants. The results demonstrate that our proposed MAF module and CNN-Transformer structure can fully utilize features and improve the performance of pothole segmentation. In addition, it also demonstrates that our MAF module is able to align two features with different types and different numbers of channels, either aligned from transformer features to convolutional features, or from convolutional features to transformer features.

Further, we select one lightweight and one heavyweight well-performing variant from both C2T and T2C variants, respectively. The results of these selected variants are displayed in Tab. II. We also test the runtime of each variant on RTX3090. The table shows that C2T-C0T0, T2C-C1T1, and T2C-C2T3 achieve similar performance, but the inference speed of C2T-C0T0 is the fastest among all the selected variants. Similarly, although the results of C2T-C0T4 are better than those of C2T-C0T0, the inference speed of C2T-C0T0 is around 2.5 times that of C2T-C0T4. For the C0T0 architecture, the superior performance of the C2T strategy over T2C may be attributed to the following reasons: In the T2C

Authorized licensed use limited to: CITY UNIV OF HONG KONG. Downloaded on November 05,2025 at 04:15:11 UTC from IEEE Xplore. Restrictions apply.

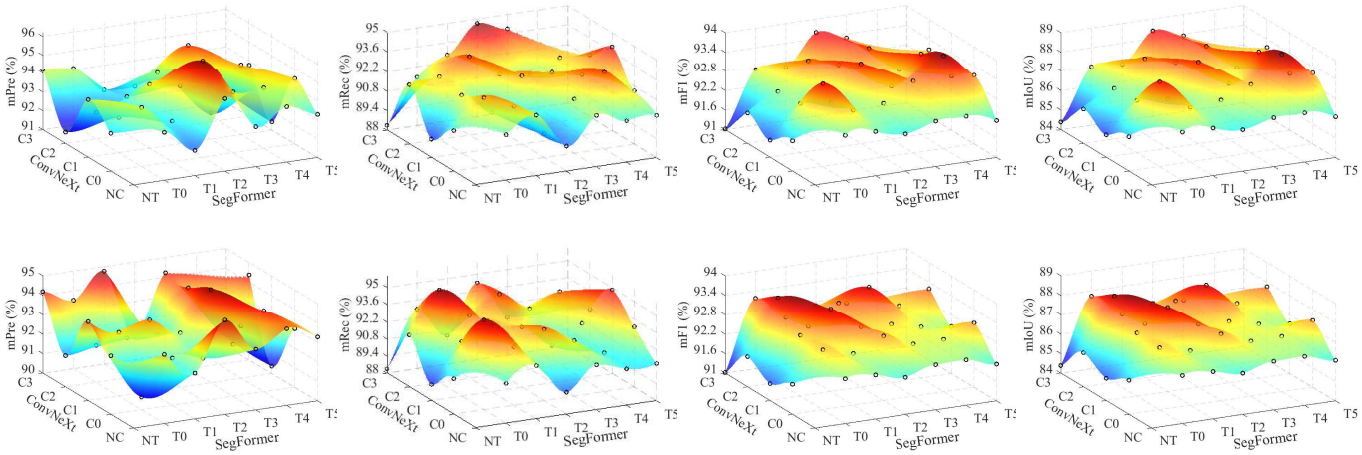


Fig. 3. The results (%) for the ablation study on encoder. The above four figures are the results of the variants with the C2T alignment strategy, and the below four figures are the results of the variants with the T2C alignment strategy. 3-D surfaces are obtained by fitting the results of different variants and visualized with the *jet* color map. The values are increased from blue to red.

structure, excessively high feature dimensions are employed to encode the fused features. However, for features extracted by lightweight encoders, the dimensionality provided by the C2T strategy is already sufficient to effectively represent the critical features for pothole segmentation. Based on the above results, we design MMFSeg as the same structure as C2T-C0T0 to achieve a trade-off between accuracy and efficiency.

2) Ablation on MAF: We conduct ablation studies on the structure of the MAF module to show the benefits of each block in our proposed MAF module. Since the alignment block is used to modify the channel for element-wise addition, we remove the other two blocks from the MAF module to design variants. The details of the variants are listed as follows:

1) *w/o* s: The spatial block is removed from the MAF module.
2) *w/o* c: The channel block is removed from the MAF module. The element-wise addition is placed at the end of the alignment block.
3) *w/o* c&s: The channel block and spatial block are removed from the MAF module. The element-wise addition is placed at the end of the alignment block.
4) *Cat*: The MAF module is replaced with a concatenation method along the channel axis.

The variants are designed based on the lightweight and well-performing C2T-C0C0 and T2C-C1T1 variants. We also compare the performance of these variants with that of single-structure variants from Section IV-D.1, which are denoted as *Conv* and *Trans*. It should be noted that the variants based on C2T-C0C0/T2C-C1C1 should be compared with variants C0NT/C1NT and NCT0/NCT1. The results of the variants are shown in Fig. 4. We can see that the mPre and mRec results for the variants are inconsistent. For example, the mPre of the *w/o* c&s variant based on C2T-C0T0 is better than that of the *Cat* variant. However, the mRec of the *w/o* c&s variant based on C2T-C0T0 is inferior to that of the *Cat* variant. However, the mF1 and mIoU for the variants are consistent. Compared to the variants *Trans*, *Cat*, and *w/o* c&s, we can find that the fusion by concatenation along the channel axis and the

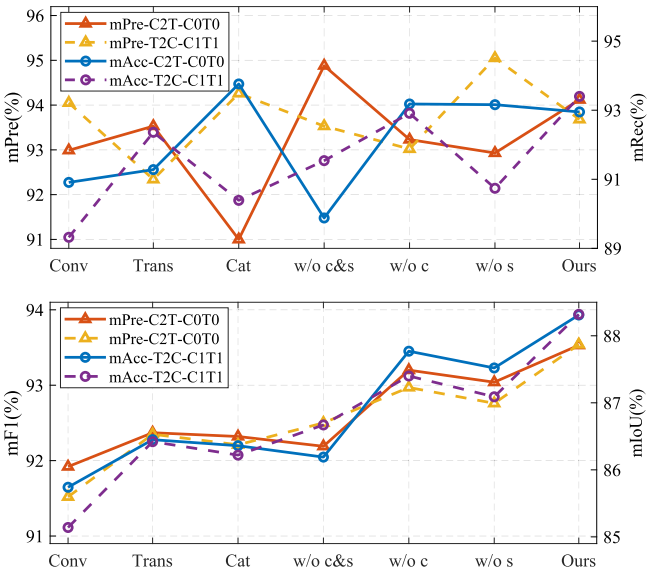


Fig. 4. The results (%) of the ablation study on the structure of the MAF module.

channel-interpolation method (alignment block) cannot well fuse the features extracted by the different-structure encoders. Specifically, the variants using these methods may achieve higher or lower results than a single structure variant. The results of the variants *w/o* c and *w/o* s are better than those of the *Trans* and *Conv*. This illustrates that our proposed channel block and spatial block in the MAF module can improve the fusion results. Compared to the variants *w/o* c&s, *w/o* c, and *w/o* s, we can find that both channel block and spatial block can boost the performance. The variants with our proposed MAF module achieve the best results among the variants adopting the same alignment strategy, which demonstrates the benefits of our MAF module.

3) Ablation on the Feature Types in MAF: We conduct ablation studies on the feature types in the MAF module to show the contributions of the CNN and Transformer features. We design some variants by removing or doubling one of features. The details of the variants are listed as follows:

Authorized licensed use limited to: CITY UNIV OF HONG KONG. Downloaded on November 05,2025 at 04:15:11 UTC from IEEE Xplore. Restrictions apply.

TABLE III

THE QUANTITATIVE RESULTS (%) ON THE POTHOLE-600 DATASET OF ALL COMPARED NETWORKS AND OUR MMFSEG. † MEANS THAT THE RESULTS ARE DIRECTLY IMPORTED FROM THE ORIGINAL PAPER [2], WHICH IS TRAINED WITH THE AUGMENTED DATASET. THE OTHER NETWORKS ARE TRAINED WITH THE TRAINING SET OF THE POTHOLE-600 DATASET. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD FONT

Network	Background				Pothole				mPre	mRec	mF1	mIoU
	Pre	Rec	F1	IoU	Pre	Rec	F1	IoU				
AARTFNet [12]	95.93	99.57	97.72	95.53	93.05	57.88	71.37	55.48	94.49	78.72	84.54	75.51
GMNet [43]	97.66	99.11	98.38	96.81	89.59	76.29	82.40	70.07	93.62	87.70	90.39	83.44
AMFNet [45]	96.13	**99.67**	97.87	95.83	**94.87**	60.02	73.52	58.13	**95.50**	79.85	85.70	76.98
InconSeg [11]	97.60	99.43	98.51	97.06	92.96	75.63	83.41	71.54	95.28	87.53	90.96	84.30
MAFNet [2]	97.77	99.32	98.54	97.12	91.89	77.40	84.03	72.45	94.83	88.36	91.28	84.78
MAFNet†[2]	98.83	98.54	98.69	97.41	85.88	88.39	87.11	77.17	92.35	93.47	92.90	87.29
CENet [60]	98.37	98.85	98.61	97.26	87.98	83.70	85.79	75.11	93.18	91.28	92.20	86.19
EAEFNet [40]	97.63	99.34	98.48	97.00	92.03	75.98	83.24	71.29	94.83	87.66	90.86	84.15
RoadSeg [38]	98.50	98.49	98.50	97.04	85.00	85.08	85.04	73.97	91.75	91.79	91.77	85.51
FRNet [39]	97.69	99.33	98.50	97.05	92.01	76.56	83.57	71.78	94.85	87.95	91.04	84.42
MMSMCNet [48]	98.23	98.92	98.57	97.19	88.39	82.25	85.21	74.23	93.31	90.58	91.89	85.71
PotCrackSeg [14]	98.08	99.29	98.69	97.40	91.95	80.66	85.94	75.34	95.02	89.98	92.31	86.37
MMFSeg (Ours)	**98.69**	98.98	**98.84**	**97.70**	89.56	**86.93**	**88.22**	**78.92**	94.12	**92.95**	**93.53**	**88.31**

TABLE IV

THE QUANTITATIVE RESULTS (%) ON THE NO-4K DATASET OF OUR MMFSEG AND ALL COMPARED NETWORKS. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD FONT

Network	Background				Pothole				mPre	mAcc	mF1	mIoU
	Pre	Acc	F1	IoU	Pre	Acc	F1	IoU				
AARTFNet [12]	99.43	98.80	99.11	98.24	81.07	90.10	85.34	74.43	90.25	94.45	92.23	86.34
GMNet [43]	99.32	99.10	99.21	98.43	84.86	88.04	86.42	76.09	92.09	93.57	92.82	87.26
AMFNet [45]	99.31	99.24	99.27	98.56	86.80	87.94	87.37	77.57	93.06	93.59	93.32	88.06
InconSeg [11]	99.36	99.22	99.29	98.59	86.68	88.81	87.73	78.14	93.02	94.01	93.51	88.37
MAFNet [2]	99.37	98.85	99.11	98.23	81.59	88.94	85.10	74.07	90.48	93.90	92.11	86.15
CENet [60]	99.38	99.01	99.19	98.40	83.65	89.12	86.30	75.90	91.51	94.06	92.75	87.15
EAEFNet [40]	99.31	99.18	99.25	98.51	86.01	87.96	86.98	76.95	92.66	93.57	93.11	87.73
RoadSeg [38]	99.37	99.12	99.24	98.50	85.24	88.92	87.04	77.06	92.30	94.02	93.14	87.78
FRNet [39]	99.40	99.01	99.21	98.42	83.78	89.57	86.58	76.34	91.59	94.29	92.89	87.38
MMSMCNet [48]	99.45	99.11	99.28	98.56	85.21	90.31	87.69	78.07	92.33	94.71	93.48	88.32
PotCrackSeg [14]	**99.49**	99.01	99.25	98.51	83.98	**91.10**	87.39	77.61	91.73	**95.05**	93.32	88.06
MMFSeg (Ours)	99.37	**99.39**	**99.38**	**98.77**	**89.29**	88.96	**89.12**	**80.38**	**94.33**	94.17	**94.25**	**89.57**

TABLE V

THE RESULTS (%) OF SELECTED VARIANTS FROM THE ABLATION STUDY ON THE FEATURE TYPES IN THE MAF MODULE

Variant	MAF Module		mPre	mAcc	mIoU	mF1
	Conv	Trans				
O-C_n	C_n	–	92.48	93.31	87.28	92.89
D-C_n	C_n	C_n	92.83	**93.33**	87.58	93.08
O-T_n	–	T_n	92.06	92.84	86.56	92.45
D-T_n	T_n	T_n	**94.21**	90.78	86.53	92.41
MMFSeg (Ours)	C_n	T_n	94.12	92.95	**88.31**	**93.53**

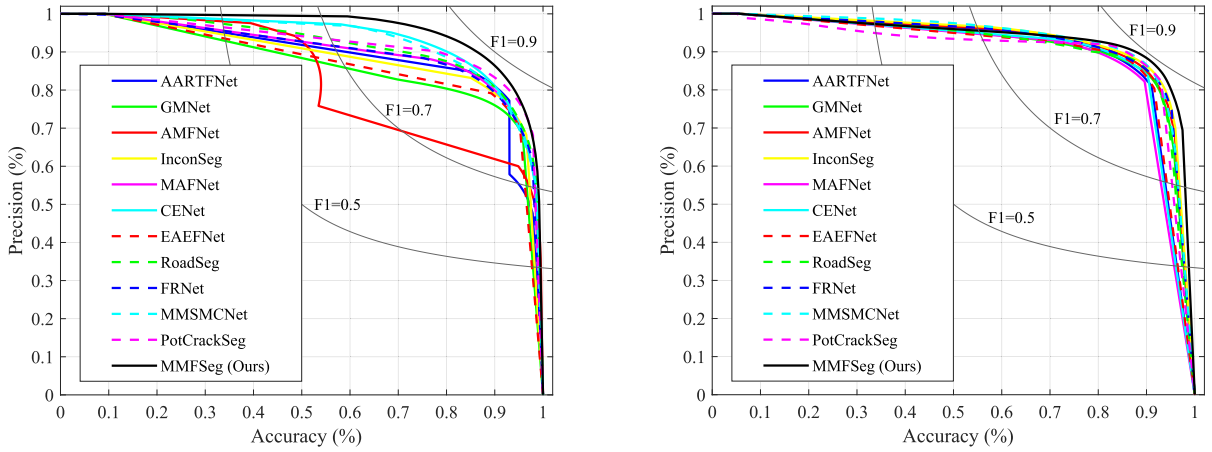
1) O-C_n: Transformer features are removed from the MAF module. The branches of the Transformer features are removed from each block of the MAF module.
2) D-C_n: Transformer features are replaced by CNN features. The number of channels in each block of the MAF module are adjusted with the input.
3) O-T_n: CNN features are removed from the MAF module. The branches of the CNN features are removed from each block of the MAF module.
4) D-T_n: CNN features are replaced by Transformer features. The number of channels in each block of the MAF module are adjusted with the input.

The experimental results are displayed in Tab. V. By comparing the results of Variant O-C_n, Variant O-T_n, and our MMFSeg, we can see that when there is only one type of feature in the MAF module, that is, only CNN features or only Transformer features, the performance is inferior to that when both types of features are present. In addition, by comparing Variant O-C_n with Variant D-C_n, and Variant O-T_n with Variant D-T_n, it can be found that simply doubling a single type of feature cannot effectively improve the performance. We also find that the variants with the CNN features outperform the variants with the Transformer feature. The reason may be that the textures of the potholes are similar to those of the background area, which imposes challenges to the self-attention mechanism in the Transformer feature to identify the pothole features. Therefore, the fused features take into account both local and global features, containing more features, and thus achieving better accuracy. The experimental results demonstrate the contribution of each type of features in the MAF module, and also the effectiveness of our proposed MAF module to fuse two different types of features.

E. Comparative Experiment

We compare our MMFSeg with some well-known networks: AARTFNet [12], GMNet [43], AMFNet [45], InconSeg [11],

Authorized licensed use limited to: CITY UNIV OF HONG KONG. Downloaded on November 05,2025 at 04:15:11 UTC from IEEE Xplore. Restrictions apply.



(a) Pothole-600 (b) NO-4K

Fig. 5. The Precision-Recall curve for our MMFSeg and all the compared networks trained with the Pothole-600 dataset and the NO-4K dataset.

MAFNet [2], CENet [60], EAEFNet [40], RoadSeg [38], FRNet [39], MMSMCNet [48], and PorCrackSeg [14]. All the networks are trained and tested with the same dataset. The four metrics (i.e., Pre, Rec, F1, and IoU) of *background* and *pothole* and the mean values of the metrics are employed to evaluate the performance of each network. We also test the inference speed of each network on the NVIDIA RTX3060 and RTX3090 graphics cards.

1) Quantitative Results on `Pothole-600`*:* The quantitative results on the `Pothole-600` of all compared networks and our MMFSeg are displayed in Tab. III. We can see that our MMFSeg achieves the best performance among all the compared networks. Specifically, for the *background* class, our MMFSeg achieves the best results in terms of Pre, F1, and IoU. For the *pothole* class, our MMFSeg achieves the best results in terms of Rec, F1, and IoU. In addition, our MMFSeg outperforms the other compared networks in terms of mRec, mF1, and mIoU. Tab. III displays our re-trained results of MAFNet trained with the training set of the `Pothole-600` dataset. The table also displays the results of MAFNet trained with the augmented dataset [2] (denoted as MAFNet†), which are directly importing from its original paper [2]. Our MMFSeg network outperforms MAFNet† in terms of mRec by 0.52%, mF1 by 0.63%, and mIoU by 1.02%. In all the compared networks trained with the training set of the `Pothole-600` dataset, PotCrackSeg achieves the second-best results. Our MMFSeg network outperforms PotCrackSeg in terms mRec by 2.97%, mF1 by 1.22%, and mIoU by 1.94%. Although AMFNet outperforms our MMFSeg in terms of Rec on the *background* class, Pre on the *pothole* class, and mPre on both the classes, the other metrics of AMFNet are significantly inferior to our MMFSeg.

We also plot the Precision-Recall (PR) curve for our MMFSeg and all the compared networks trained with the training set of the `Pothole-600` dataset. Fig. 5 (a) shows the PR curve. We can see that the curve of our MMFSeg is smoother and higher than that of the compared networks. This illustrates that the precision of our MMFSeg is higher than the compared networks for the same recall. Overall, the results in Tab. III and Fig. 5 (a) demonstrate the superiority of our proposed MMFSeg.

TABLE VI
THE EFFICIENCY ON NVIDIA RTX3060 AND RTX3090 GRAPHICS CARDS

Network	Pothole-600				NO-4K			
	3060		3090		3060		3090	
	ms	FPS	ms	FPS	ms	FPS	ms	FPS
AARTFNet [12]	58.7	17.0	25.3	39.5	127.1	7.9	42.4	23.6
GMNet [43]	58.6	17.1	29.1	34.4	118.1	8.5	49.3	20.3
AMFNet [45]	87.5	11.4	52.8	18.9	195.9	5.1	72.6	13.8
InconSeg [11]	101.9	9.8	46.0	21.8	203.2	4.9	74.7	13.4
MAFNet [2]	26.3	38.0	18.9	53.1	73.1	13.7	29.0	34.5
CENet [60]	103.0	9.7	51.4	19.5	243.8	4.1	88.8	11.3
EAEFNet [40]	67.3	14.9	37.4	26.8	134.7	7.4	58.3	17.2
RoadSeg [38]	146.4	6.8	69.2	14.5	313.1	3.2	117.9	8.5
FRNet [39]	41.5	24.1	22.6	44.2	82.1	12.2	34.9	28.7
MMSMCNet [48]	113.6	8.8	64.0	15.6	280.4	3.6	107.3	9.3
PotCrackSeg [14]	66.8	15.0	31.4	31.9	177.3	5.6	59.4	16.8
MMFSeg (Ours)	52.3	19.1	28.9	34.6	124.2	8.1	48.7	20.5

2) Quantitative Results on `NO-4K`*:* The quantitative results on the `NO-4K` dataset are displayed in Tab. IV. We can see that our MMFSeg achieves the best performance among all the compared networks, for example, our MMFSeg achieves the best results in terms of F1 and IoU on the *background* and *pothole* classes. Our MMFSeg outperforms the second-best MMSMCNet in terms mF1 by 0.74% and mIoU by 1.25%. We also plot the Precision-Recall (PR) curve for our MMFSeg and all the compared networks trained with the `NO-4K` dataset. Fig. 5 (b) shows the PR curve. We can see that the curve of our MMFSeg is smoother and higher than that of the compared networks at high accuracy and high F1 values. This illustrates that the precision of our MMFSeg is higher than that of the compared networks for the same recall. The results in Tab. IV and Fig. 5 (b) show the superiority of our proposed MMFSeg on the `NO-4K` dataset.

Overall, the results on the `Pothole-600` dataset and `NO-4K` dataset demonstrate the generalization capabilities of our proposed MMFSeg.

Authorized licensed use limited to: CITY UNIV OF HONG KONG. Downloaded on November 05,2025 at 04:15:11 UTC from IEEE Xplore. Restrictions apply.

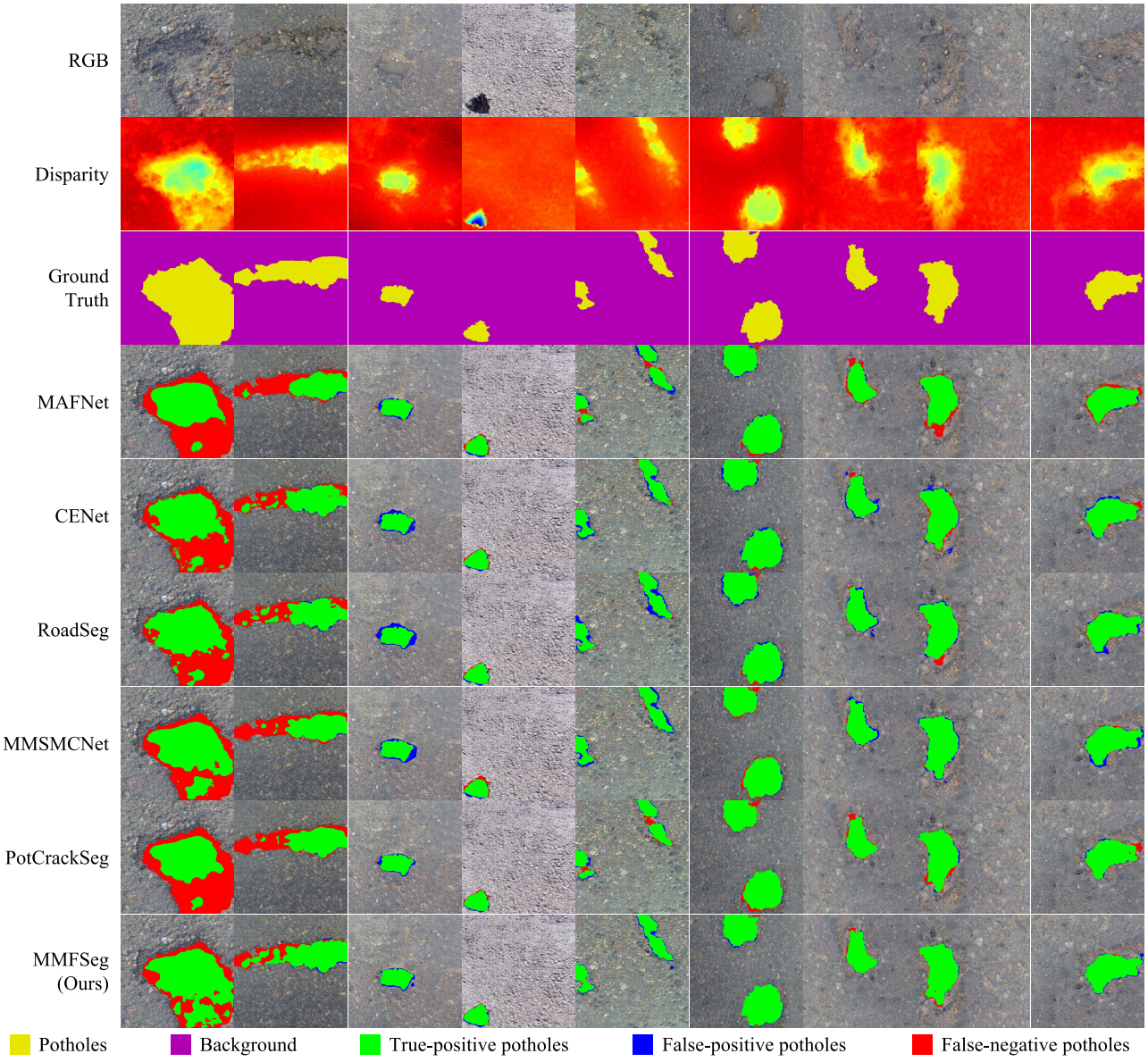


Fig. 6. Sample qualitative demonstrations for our MMFSeg and the compared networks with the top five performance trained with the `Pothole-600` dataset. The segmentation results are overlaid on the RGB images with different colors. The first seven columns show different pothole images. The last two columns are different views of the pothole contained in 7-th columns. The results show that our network performs better at the edges of potholes.

3) The Efficiency: We also evaluate the efficiency of the networks on NVIDIA RTX3060 and RTX3090 graphics cards. The results are displayed in Tab. VI. For the `Pothole-600` dataset with the 512×512 resolution, our MMFSeg achieves nearly real-time speed (19.1 FPS) on RTX3060 and real-time speed (34.6 FPS) on RTX3090. Compared with the other networks, we can find that our proposed MMFSeg achieves the third place in inference speed on RTX3060, and the fourth place in inference speed on RTX3090. For the `NO-4K` dataset with the 576×1024 resolution, the inference speed of our MMFSeg also ranks among the top four of all the networks, and MMFSeg still achieves nearly real-time inference (20.5 FPS) on RTX3090. So, our MMFSeg achieves a trade-off between accuracy and efficiency compared with the other networks.

4) Qualitative Demonstrations on `Pothole-600`: Fig. 6 shows sample qualitative demonstrations for our MMFSeg and the top five performance networks trained with the training set of the `Pothole-600` dataset from Tab. III. The first three rows are RGB images, disparity images, and the ground truth. The other rows are the segmentation results of the networks. We overlay the segmentation results with different colors on the RGB images.

The images of the first two columns contain large potholes. One common feature of these two potholes is the significant difference in depth within them. For example, in the disparity image of the first column, the depth of the upper left side of the pothole is significantly deeper than the lower right side. From the segmentation results, we can find that the difference in depth increases the difficulty of segmenting the regions

Authorized licensed use limited to: CITY UNIV OF HONG KONG. Downloaded on November 05,2025 at 04:15:11 UTC from IEEE Xplore. Restrictions apply.

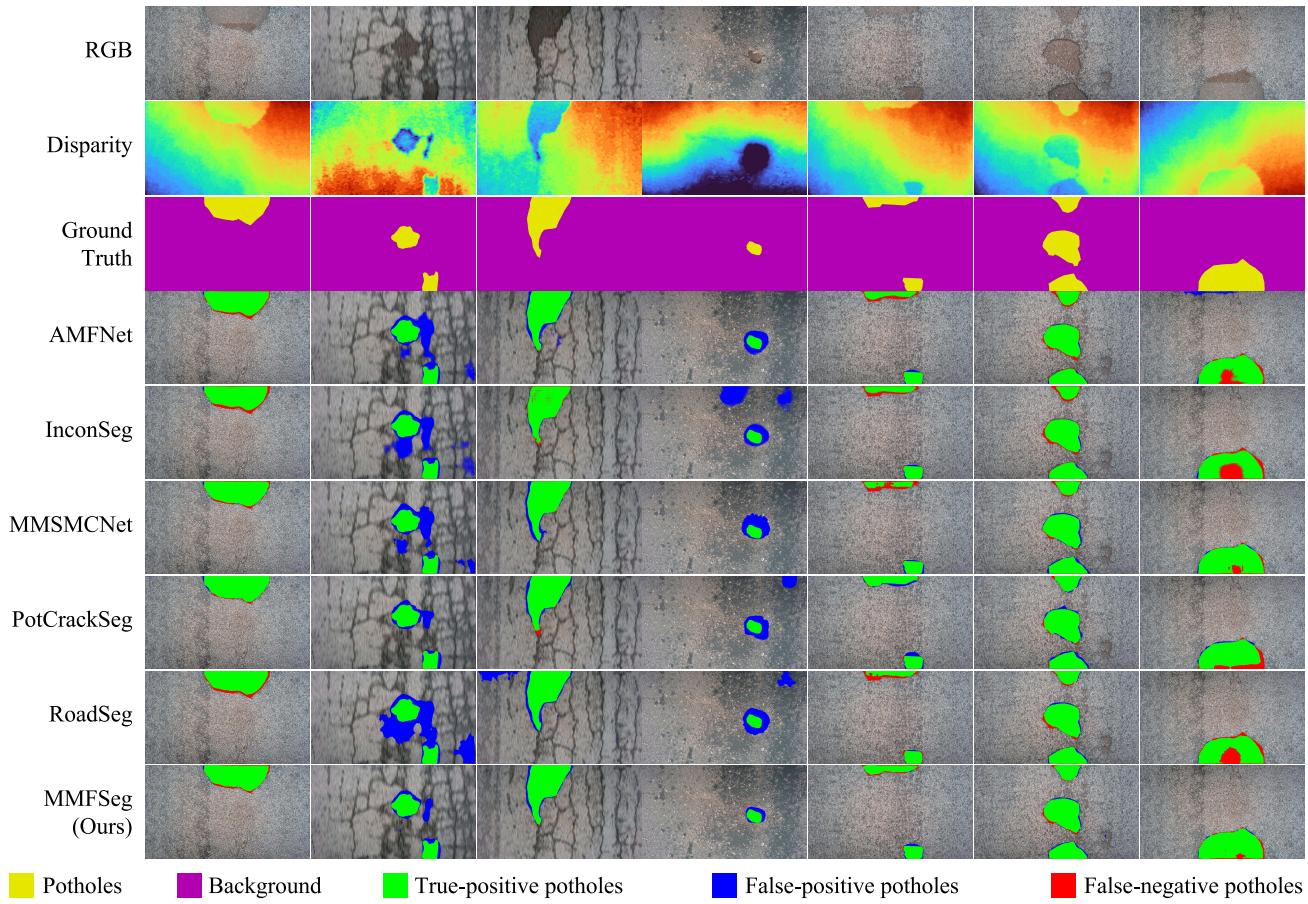


Fig. 7. Sample qualitative demonstrations for our MMFSeg and the compared networks with top five performance trained with the NO-4K dataset. The segmentation results are overlaid on the RGB images with different colors.

with shallower depth in potholes. The figure shows the five compared networks incorrectly segment the shallower-depth regions. In contrast, our network achieves better results in these regions, which correctly segments most of the potholes.

The 3-rd and 4-th columns show images containing small potholes. The results show that all the networks correctly segment most regions of the potholes. The main segmentation errors occur at the edges of the potholes. In contrast, our MMFSeg is closer to the ground truth at the edge of the pothole. Each image of the fifth and sixth columns contains two potholes. We can see that our network achieves better results within these areas (i.e., the center area of the potholes in the fifth column.) and sharp areas of potholes (i.e., the upper left area of the above pothole in the sixth column).

The last three columns show the same pothole in different viewpoint images obtained with different augmentations. The results illustrate that our network achieves better results for the same pothole at different viewpoints. Overall, our network is significantly superior to the well-known networks.

5) Qualitative Demonstrations on NO-4K: Some sample qualitative demonstrations for our MMFSeg and the top five performance networks trained with the NO-4K dataset are shown in Fig. 7. The first column and the third column demonstrate that our MMFSeg outperforms the other networks on the edges of the potholes. In the second column, the wet areas interfere with the segmentation of the potholes. In contrast, our MMFSeg is able to overcome the interferences from the wet areas and achieve high segmentation accuracy. The fourth column shows the high accuracy of our MMFSeg for small potholes. The first column and the last column show the same pothole in different viewpoints obtained with data augmentation. The results also show that our network achieves better results even with different viewpoints for the same pothole. In addition, our MMFSeg could overcome the influences of the puddled water in the potholes, achieving the best performance.

F. Main Findings From Experiments

From the experimental results, we have the following five main findings:

1) Our proposed MAF module can fuse the two types of features (i.e., CNN features and transformer features) to achieve superiority over those using a single-type feature.
2) Our proposed MMFSeg achieves the optimal accuracy compared with well-known networks on two public datasets (Pothole-600 and NO-4K).
3) Experiments show that our MMFSeg exhibits better generalization ability than the well-known networks.
4) Our proposed MMFSeg achieves a trade-off between accuracy and efficiency.

Authorized licensed use limited to: CITY UNIV OF HONG KONG. Downloaded on November 05,2025 at 04:15:11 UTC from IEEE Xplore. Restrictions apply.

5) Our proposed MMFSeg outperforms the other well-known networks at the edges of potholes.

V. Conclusion and Future Work

We proposed MMFSeg with a Transformer-convolution encoder for road pothole segmentation. We use a Transformer-based encoder and a convolution-based encoder to extract features from the two modalities of data. We also designed the late-fusion MAF module to align and fuse the features extracted by encoders that have different structures and different numbers of channels. Our proposed methods are able to better fuse different types of features and improve the performance of road-pothole segmentation. The experimental results demonstrate the superiority of our MMFSeg in terms of accuracy and efficiency compared with the well-known networks. Specifically, our MMFSeg achieves 1.02% improvement on the `Pothole-600` dataset and 1.25% improvement on the `NO-4K` dataset in terms of mIoU.

Although our MMFSeg is superior to the other networks, some limitations still need to be addressed. For example, the accuracy of the edge area of potholes needs to be improved. A potential solution is to introduce edge information to improve the performance at the edge areas of potholes. Specifically, we can refine the decoder architecture to output edge information as supervisory information, and introduce an edge-aware loss function to improve the performance of the edge areas. In addition, our MMFSeg relies on RGB-D data for segmentation, so in scenarios where depth data are unavailable, the network may not be able to work. A potential solution is to use the disparity data as privileged information, and through the distillation method, transfer the ability of MMFSeg to an RGB-only network to segment potholes.

References

[1] R. Hafezzadeh, F. Autelitano, and F. Giuliani, "Asphalt-based cold patches for repairing road potholes—An overview," *Construct. Building Mater.*, vol. 306, Nov. 2021, Art. no. 124870.

[2] Z. Feng et al., "MAFNet: Segmentation of road potholes with multi-modal attention fusion network for autonomous vehicles," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–12, 2022.

[3] T. Yin, W. Zhang, J. Kou, and N. Liu, "Promoting automatic detection of road damage: A high-resolution dataset, a new approach, and a new evaluation criterion," *IEEE Trans. Autom. Sci. Eng.*, vol. 22, pp. 2472–2484, 2025.

[4] H. Ghahremannezhad, H. Shi, and C. Liu, "Object detection in traffic videos: A survey," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 7, pp. 6780–6799, Jul. 2023.

[5] Y. Feng, Z. Feng, W. Hua, and Y. Sun, "Multimodal-XAD: Explainable autonomous driving based on multimodal environment descriptions," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 12, pp. 19469–19481, Dec. 2024.

[6] H. Li, H. K. Chu, and Y. Sun, "Improving RGB-thermal semantic scene understanding with synthetic data augmentation for autonomous driving," *IEEE Robot. Autom. Lett.*, vol. 10, no. 5, pp. 4452–4459, May 2025.

[7] S. D. Nguyen, T. S. Tran, V. P. Tran, H. J. Lee, M. J. Piran, and V. P. Le, "Deep learning-based crack detection: A survey," *Int. J. Pavement Res. Technol.*, vol. 16, no. 4, pp. 943–967, 2022.

[8] R. Fan, H. Wang, Y. Wang, M. Liu, and I. Pitas, "Graph attention layer evolves semantic segmentation for road pothole detection: A benchmark and algorithms," *IEEE Trans. Image Process.*, vol. 30, pp. 8144–8154, 2021.

[9] J. Dong et al., "CBAM-optimized automatic segmentation and reconstruction system for monocular images with asphalt pavement potholes," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 8, pp. 1–18, Aug. 2024.

[10] Aparna, Y. Bhatia, R. Rai, V. Gupta, N. Aggarwal, and A. Akula, "Convolutional neural networks based potholes detection using thermal imaging," *J. King Saud Univ.-Comput. Inf. Sci.*, vol. 34, no. 3, pp. 578–588, Mar. 2022.

[11] Z. Feng, Y. Guo, D. Navarro-Alarcon, Y. Lyu, and Y. Sun, "InconSeg: Residual-guided fusion with inconsistent multi-modal data for negative and positive road obstacles segmentation," *IEEE Robot. Autom. Lett.*, vol. 8, no. 8, pp. 4871–4878, Aug. 2023.

[12] R. Fan, H. Wang, M. J. Bocus, and M. Liu, "We learn better road pothole detection: From attention aggregation to adversarial domain adaptation," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 285–300.

[13] X. Li and J. Zhou, "MASNet: Road semantic segmentation based on multiscale modality fusion perception," *IEEE Trans. Instrum. Meas.*, vol. 73, pp. 1–13, 2024.

[14] 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. Vehicles*, vol. 9, no. 4, pp. 4687–4697, Apr. 2024.

[15] Y. Sun, W. Zuo, P. Yun, H. Wang, and M. Liu, "FuseSeg: Semantic segmentation of urban scenes based on RGB and thermal data fusion," *IEEE Trans. Autom. Sci. Eng.*, vol. 18, no. 3, pp. 1000–1011, Jul. 2021.

[16] W. Zhou, J. Yang, W. Yan, and M. Fang, "RDNet-KD: Recursive encoder, bimodal screening fusion, and knowledge distillation network for rail defect detection," *IEEE Trans. Autom. Sci. Eng.*, vol. 22, pp. 2031–2040, 2025.

[17] S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz, and D. Terzopoulos, "Image segmentation using deep learning: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 7, pp. 3523–3542, Jul. 2022.

[18] H. Thisanke, C. Deshan, K. Chamith, S. Seneviratne, R. Vidanaarachchi, and D. Herath, "Semantic segmentation using vision transformers: A survey," *Eng. Appl. Artif. Intell.*, vol. 126, Nov. 2023, Art. no. 106669.

[19] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. 18th Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, vol. 9351, Munich, Germany. Cham, Switzerland: Springer, 2015, pp. 234–241.

[20] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. 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. 2017.

[21] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder–decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 801–818.

[22] Z. Han, M. Jian, and G.-G. Wang, "ConvUNeXt: An efficient convolution neural network for medical image segmentation," *Knowl.-Based Syst.*, vol. 253, Oct. 2022, Art. no. 109512.

[23] 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.*, Jun. 2022, pp. 11976–11986.

[24] J. Feng et al., "Robotic inspection and data analytics to localize and visualize the structural defects of concrete infrastructure," *IEEE Trans. Autom. Sci. Eng.*, early access, Jan. 27, 2025, doi: 10.1109/TASE.2025.3535227.

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

[26] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent.*, 2020.

[27] H. Wu, Z. Zhao, and Z. Wang, "META-UNet: Multi-scale efficient transformer attention unet for fast and high-accuracy polyp segmentation," *IEEE Trans. Autom. Sci. Eng.*, vol. 21, no. 3, pp. 4117–4128, Jul. 2024.

[28] Y. Liu et al., "A survey of visual transformers," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 6, pp. 7478–7498, Jun. 2023.

[29] E. Xie et al., "SegFormer: Simple and efficient design for semantic segmentation with transformers," in *Proc. Adv. Neural Inf. Process. Sys.*, vol. 34, Dec. 2021, pp. 12077–12090.

[30] Q. Wan, Z. Huang, J. Lu, G. Yu, and L. Zhang, "SeaFormer: Squeeze-enhanced axial transformer for mobile semantic segmentation," in *Proc. 11th Int. Conf. Learn. Represent.*, 2023.

[31] H. Du, J. Wang, M. Liu, Y. Wang, and E. Meijering, "SwinPA-Net: Swin transformer-based multiscale feature pyramid aggregation network for medical image segmentation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 4, pp. 5355–5366, Apr. 2024.

Authorized licensed use limited to: CITY UNIV OF HONG KONG. Downloaded on November 05,2025 at 04:15:11 UTC from IEEE Xplore. Restrictions apply.

[32] B. Chen, Y. Liu, Z. Zhang, G. Lu, and A. W. K. Kong, "TransAttUnet: Multi-level attention-guided U-Net with transformer for medical image segmentation," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 8, no. 1, pp. 55–68, Feb. 2024.

[33] X. Guo et al., "MorFormer: Morphology-aware transformer for generalized pavement crack segmentation," *IEEE Trans. Intell. Transp. Syst.*, vol. 26, no. 6, pp. 8219–8232, Jun. 2025.

[34] V. Likhosherstov, K. Choromanski, and A. Weller, "On the expressive flexibility of self-attention matrices," in *Proc. AAAI Conf. Artif. Intell.*, Jun. 2023, vol. 37, no. 7, pp. 8773–8781.

[35] D. Ramachandram and G. W. Taylor, "Deep multimodal learning: A survey on recent advances and trends," *IEEE Signal Process. Mag.*, vol. 34, no. 6, pp. 96–108, Nov. 2017.

[36] T. Baltrusaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 423–443, Feb. 2019.

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

[38] R. Fan, H. Wang, P. Cai, and M. Liu, "Sne-roadseg: Incorporating surface normal information into semantic segmentation for accurate freespace detection," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 340–356.

[39] W. Zhou, E. Yang, J. Lei, and L. Yu, "FRNet: Feature reconstruction network for RGB-D indoor scene parsing," *IEEE J. Sel. Topics Signal Process.*, vol. 16, no. 4, pp. 677–687, Jun. 2022.

[40] M. Liang, J. Hu, C. Bao, H. Feng, F. Deng, and T. L. Lam, "Explicit attention-enhanced fusion for RGB-thermal perception tasks," *IEEE Robot. Autom. Lett.*, vol. 8, no. 7, pp. 4060–4067, Jul. 2023.

[41] Y. Cai, W. Zhou, L. Zhang, L. Yu, and T. Luo, "DHFNet: Dual-decoding hierarchical fusion network for RGB-thermal semantic segmentation," *Vis. Comput.*, vol. 40, no. 1, pp. 169–179, Jan. 2024.

[42] Y. Zhang, W. Zhou, X. Ran, and M. Fang, "Lightweight dual stream network with knowledge distillation for RGB-D scene parsing," *IEEE Signal Process. Lett.*, vol. 31, pp. 855–859, 2024.

[43] W. Zhou, J. Liu, J. Lei, L. Yu, and J.-N. Hwang, "GMNet: Graded-feature multilabel-learning network for RGB-thermal urban scene semantic segmentation," *IEEE Trans. Image Process.*, vol. 30, pp. 7790–7802, 2021.

[44] W. Zhou, F. Sun, and W. Qiu, "MSNet: Multiple strategy network with bidirectional fusion for detecting salient objects in RGB-D images," *IEEE Trans. Autom. Sci. Eng.*, vol. 22, pp. 4341–4353, 2025.

[45] Z. Feng, Y. Feng, Y. Guo, and Y. Sun, "Adaptive-mask fusion network for segmentation of drivable road and negative obstacle with untrustworthy features," in *Proc. IEEE Intell. Vehicles Symp.*, Jun. 2023, pp. 1–6.

[46] W. Zhou, H. Zhang, Y. Liu, and T. Luo, "Enhancing RGB-D mirror segmentation with a neighborhood-matching and demand-modal adaptive network using knowledge distillation," *IEEE Trans. Autom. Sci. Eng.*, vol. 22, pp. 12679–12692, 2025.

[47] Y. Feng et al., "SNE-RoadSegV2: Advancing heterogeneous feature fusion and fallibility awareness for freespace detection," *IEEE Trans. Instrum. Meas.*, vol. 74, pp. 1–9, 2025.

[48] W. Zhou, H. Zhang, W. Yan, and W. Lin, "MMSMCNet: Modal memory sharing and morphological complementary networks for RGB-T urban scene semantic segmentation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 12, pp. 7096–7108, Dec. 2023.

[49] W. Zhou, B. Jian, and Y. Liu, "Feature contrast difference and enhanced network for RGB-D indoor scene classification in Internet of Things," *IEEE Internet Things J.*, vol. 12, no. 11, pp. 17610–17621, Jun. 2025.

[50] 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. Vehicles*, vol. 9, no. 7, pp. 5163–5172, Jul. 2024.

[51] J. Huang et al., "RoadFormer+: Delivering RGB-X scene parsing through scale-aware information decoupling and advanced heterogeneous feature fusion," *IEEE Trans. Intell. Vehicles*, vol. 10, no. 5, pp. 1–10, May 2025.

[52] X. Han, C. Nguyen, S. You, and J. Lu, "Single image water hazard detection using fcn with reflection attention units," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 105–120.

[53] X. Huang et al., "LEPS: A lightweight and effective single-stage detector for pothole segmentation," *IEEE Sensors J.*, vol. 24, no. 14, pp. 22045–22055, Jul. 2024.

[54] Q. Zhou, Z. Qu, and F.-R. Ju, "A lightweight network for crack detection with split exchange convolution and multi-scale features fusion," *IEEE Trans. Intell. Vehicles*, vol. 8, no. 3, pp. 2296–2306, Mar. 2023.

[55] P. Liu, J. Yuan, and S. Chen, "A road damage segmentation method for complex environment based on improved UNet," in *Proc. Int. Conf. Image Graph.*, 2023, pp. 332–343.

[56] B. Xu, W. Shao, and X. Dong, "Drone-based wall crack detection using model-agnostic meta-learning," *IEEE Trans. Autom. Sci. Eng.*, vol. 22, pp. 15116–15128, 2025.

[57] N. Ma, R. Fan, and L. Xie, "UP-CrackNet: Unsupervised pixel-wise road crack detection via adversarial image restoration," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 10, pp. 13926–13936, Oct. 2024.

[58] Z. Feng, Y. Guo, and Y. Sun, "CPKD: Channel and position-wise knowledge distillation for segmentation of road negative obstacles," in *Proc. IEEE 26th Int. Conf. Intell. Transp. Syst. (ITSC)*, Sep. 2023, pp. 3110–3115.

[59] 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.

[60] Z. Feng, Y. Guo, and Y. Sun, "CEKD: Cross-modal edge-privileged knowledge distillation for semantic scene understanding using only thermal images," *IEEE Robot. Autom. Lett.*, vol. 8, no. 4, pp. 2205–2212, Apr. 2023.

Zhen Feng received the B.S., M.S., and Ph.D. degrees from Harbin Institute of Technology, Harbin, China, in 2017, 2019, and 2023, respectively, and the Ph.D. degree from The Hong Kong Polytechnic University, Hong Kong, in 2024.

He is currently a Post-Doctoral Fellow with the Department of Mechanical Engineering, City University of Hong Kong, Hong Kong. His current research interests include computer vision, autonomous driving, and embodied AI.

Yanning Guo received the M.S. and Ph.D. degrees in control science and engineering from Harbin Institute of Technology, Harbin, China, in 2008 and 2012, respectively.

He is currently a Professor with the Department of Control Science and Engineering, Harbin Institute of Technology, and also teaches and performs research in the fields of deep space exploration, satellite attitude control, and nonlinear control.

Rui Fan received the B.Eng. degree in automation from Harbin Institute of Technology in 2015 and the Ph.D. degree in electrical and electronic engineering from the University of Bristol in 2018.

He is currently a Full Professor with the College of Electronic and Information Engineering, Tongji University. His research interests include computer vision, deep learning, and robotics, with a specific focus on humanoid visual perception under the two-streams hypothesis.

Yuxiang Sun received the bachelor's degree from Hefei University of Technology, Hefei, China, in 2009, the master's degree from the University of Science and Technology of China, Hefei, in 2012, and the Ph.D. degree from The Chinese University of Hong Kong, Hong Kong, in 2017.

He is currently an Assistant Professor with the Department of Mechanical Engineering, City University of Hong Kong, Hong Kong. His current research interests include robotics, autonomous driving, and embodied AI.

Authorized licensed use limited to: CITY UNIV OF HONG KONG. Downloaded on November 05,2025 at 04:15:11 UTC from IEEE Xplore. Restrictions apply.