Infrared Polarization-Empowered Full-Time Road Detection via Lightweight Multi-Pathway Collaborative 2D/3D Convolutional Networks

Xueqiang Fan, Bing Lin, and Zhongyi Guo

***Abstract*— Automatic roads detection is an essential task for traffic safety and intelligent transportation systems. Recently the long-wave infrared (LWIR) polarization imaging-based road detection technique has obtained significant progresses. However, the joint analysis among multiple polarization characteristics, sparse inter-channel information (along the z-axis), and dense intra-channel information (inside the x-y plane), have not been considered effectively, hindering the effective detection of many road areas. Additionally, most of the existing methods often encounter a challenging trade-off between achieving high precision and maintaining a lightweight design. To tackle these issues, this paper presents a novel Lightweight Multi-Pathway Collaborative 2D/3D Convolutional Networks (LMPC2D3DCNet) with a small number of parameters for full-time road detection. Our LMPC2D3DCNet is the first attempt to incorporate 2D and 3D convolutional networks to balance extraction for sparse inter-channel polarization information and dense intra-channel polarization information, in which a new Cross 2D-3D Non-Local Attention (C2D3DNLA) network is proposed to derive respective latent features by exploiting both local and global polarization correlations. Meanwhile, it also follows the design of a multipath network structure that elegantly fuses plenty of low-frequency, high-frequency, and multiscale polarization information, thus obtaining more accurate modeling for road regions. Extensive experiments on one public infrared polarization dataset of road scenes demonstrate that our proposed LMPC2D3DCNet (The code will release soon on https://github.com/XueqiangF) achieves *PRE* of 96.96%, *REC* of 96.71%, *OA* of 99.45%, *F1* of 96.72, *BER* of 1.80% and *IoU* of 93.85%, and outperforms significantly state-of-the-art methods.**

***Index Terms*— Road detection, polarimetric characteristics, LWIR, attention module, collaborative 2D/3D convolutional networks.**

I. Introduction

With rapid developments of long-wave infrared (LWIR) polarization imaging technology, LWIR polarization images (LWIR-PIs) are great significance in wide-ranging application areas, such as 3D reconstruction and anti-interference object detection [1], [2], [3]. Automatic road detection from LWIR-PIs has attracted much research interest from both academics and industries in various domains, ranging from LiDAR [4] to advanced driver assistant system [5], [6], and even autonomous driving [7].

The roads of LWIR-PIs exhibit the following typical characteristics [8], [9], [10]: **i**) Context characteristics: roads, cars, roadside infrastructures, and street trees constitute the local context features, while the surrounding buildings form the global context features; **ii**) Geometric characteristics: the road shape is first wide and then narrow from the camera's perspective; and **iii**) Polarization characteristics: roads and backgrounds have different discriminative polarization information, which is the most crucial point. Recognizing these characteristics, until now only a few LWIR polarization-assisted methods have been developed to enhance the accuracy of road detection. The existing LWIR-PIs based road detection methods can be roughly grouped into the following two categories according to their work modes: *Polarization Physical Empirical Knowledge based Models (PPEKM)* and *Statistical Learning based Models (SLM)*. The *PPEKM* generally focuses on the accurate estimation of polarimetric measurements and parameters of interest, and there are only two methods, *i.e.*, the zero-distribution of angle of polarization (AoP) [11] and PCRL (Polarization characteristics of the road in LWIR) [9]. In PCRL, the performance of road pattern detection is enhanced by integrating intensity information and temporal information, as well as leveraging the distinct polarization feature between the road area and the ground in LWIR-PIs. The limitations of these methods lie in their inability to leverage the rich high-level information embedded within the LWIR images and dataset. On the contrast, the *SLM* mainly exploits Deep Convolutional Neural Networks (DCNN) that have powerful feature extraction and feature learning capabilities, *i.e.*, PolarNet [12], which has become the cutting-edge model with state-of-the-art performance. However, the PolarNet has not consider effectively the correlation among multiple polarization characteristics.

Each of the aforementioned methods possess their own distinct advantages, and do play a role in stimulating the development of this important area. Nevertheless, it still remains a huge challenge for reliable road detection from

Manuscript received 26 September 2023; revised 3 February 2024; accepted 26 March 2024. This work was supported by the National Natural Science Foundation of China under Grant 61775050. The Associate Editor for this article was K. Wang. *(Corresponding author: Zhongyi Guo.)*

The authors are with the School of Computer and Information, Hefei University of Technology, Hefei 230009, China (e-mail: fanxueqiang2022b@163.com; linbing2021s@163.com; guozhongyi@hfut.edu.cn).

Digital Object Identifier 10.1109/TITS.2024.3383405

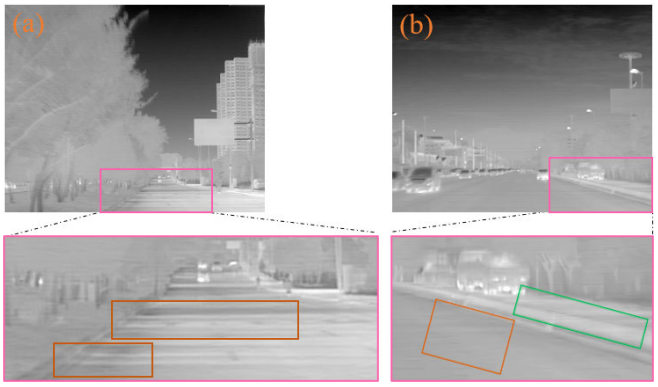

Fig. 1. Typical challenges in road detection from the LWIR-PIs. (a) road discontinuities are induced due to shadow phenomena and the occlusion of trees. (b) the similarity between a road and its surrounding ground leads to missing detection and misclassification.

LWIR-PIs, including the following tricky issues: **i**) Road detection is a standard imbalanced-learning problem since the road often takes up a small proportion in a road image; **ii**) LWIR-PIs have the characteristics of low resolution and time-varying (the gray levels of LWIR-PIs exhibit variations between daytime and nighttime LWIR-PIs); **iii**) Road detection models cannot achieve the best performances in terms of high-accuracy and lightweight; and **iv**) Surrounding objects on the roadside interfere with road recognition, for instance vehicles on the road and shadows of vegetation or buildings on the roadside. As depicted in Fig. 1, one of the usual troubles in road detection is associated with tree shadows leading to road discontinuity (as can be seen in the first example). Another challenge is that the similarity between a road and its surrounding ground leads to missed detection and wrong classification (as shown in the second example).

To cope with these issues, we present a customized LWIR-PIs based road detection model termed Lightweight Multi-Pathway Collaborative 2D/3D Convolutional Networks (LMPC2D3DCNet) with the purpose of improving the performance of road detection. The design strategies in our method are: i) mining and fusing plenty of low-frequency, high-frequency and multiscale polarization information for obtaining more accurate modeling for road regions; ii) jointing analysis among multiple polarization characteristics, sparse inter-channel information (along the $z-$axis), and dense intra-channel information (inside the $x-y$ plane); and iii) distilling latent local and global context features.

To sum up, the main contributions of this work as follows:

- We analyze the inherent deficiencies of currently automatic road detection from LWIR-PIs, and elaborately design the multi-path network architecture of LMPC2D3DCNet to capture the low-frequency, high-frequency and multiscale physical polarization coherence. Also, it has much fewer parameters than the existing road detection methods.
- To the best of our knowledge, it is the first endeavor to incorporate 2D and 3D convolutional networks to balance extraction for sparse inter-channel information and dense intra-channel polarization information embedded in multidimensional polarization characteristics.
- A new cross 2D-3D non-local attention (C2D3DNLA) network is proposed to derive respective latent features by exploiting both local and global polarization correlations.
- We carry out extensive experiments on a public infrared polarization dataset of road scenes. The experimental results demonstrate that our proposed LMPC2D3DCNet can enhance the performance of road detection maintaining a competitive inferring time, and achieves state-of-the-art performance.

The remaining of this paper is organized as follows: *Sec.* II summarizes the related works of road detection from LWIR-PIs. The details of our proposed LMPC2D3DCNet are introduced in *Sec.* III. In *Sec.* IV, dataset, evaluation metrics, and implementation details are provided, and numerous experiments are carried out to evaluate the performance of our method for road detection with LWIR-PIs. Conclusion is presented in *Sec.* V.

II. Related Work

In this section, we will introduce a concise review of road detection using LWIR-PIs and focus on the specific methods that are most relevant to our work.

A. Road Detection Models With LWIR-PI

In the past decade, numerous works (*e.g.*, ENet [13], DenseASPP [14], SegNet [15], DLT-Net [16], and DABNet [17]) have been conducted out on real-time road detection using conventional RGB cameras that provide high-resolution intensity, color, and texture information. For instance, Lu et al. [18] used a likelihood ratio classifier to re-label each pixel of input image and implemented a self-learner statistics model; Qian et al. [16] developed a unified neural network DLT-Net to detect drivable areas, lane lines and traffic objects based on Full Convolutional Networks (FCN); and Li et al. [17] employed the combination of dilated CNN and depth-wise separable CNN to design a Depth-wise Asymmetric Bottleneck (DAB) for extracting local and contextual information. The aforementioned technologies have proved to be highly effective under normal lighting conditions. However, their effectiveness will be significantly reduced in low light environments, especially at nighttime.

To overcome this limitation, a few LWIR polarization-assisted road detection methods have been proposed recently. Unlike the intensity and spectrum information used in conventional vision systems, polarization provides not only the light intensity distribution of the scene, but also the polarization feature distribution that reveals characteristic information of the object such as surface smoothness, three-dimensional (3D) normal [19], and material composition [9]. The study of full-time road detection from LWIR-PIs is still in its infancy. To our best knowledge, there are three methods (*i.e.*, the zero-distribution of AoP) [11], PCRL [9], and PolarNet [12]) available for full-time road detection. In PCRL, the distinct polarization characteristics in LWIR between the road region and the ground, combined with

the intensity and temporal information are utilized to detect roadway patterns. Further, PCRL has also constructed the first LWIR DoFP Dataset of Road Scene (LDDRS), filling a gap in this field. To fully exploit polarization information, Li et al. [12] employ a combination of FCN and spatial attention to proposed PolarNet, a two-branch network designed for road detection. Although much progress has been made, there is still room to improve the performance of road detection. A natural idea is to explore the optimization of effective multiple polarization characteristics learning for maximizing the distillation and fusion of low-frequency, high-frequency, and multiscale polarization information. Our approach is designed towards the goal of learning multiple polarization characteristics through the multi-pathway structure networks.

B. Road Semantic Segmentation Models

In recent years, deep learning techniques have started to emerge as an alternative approach to road extraction problems. Numerous researchers have performed the road extraction task as a pixel-level classification problem by using FCN, and developed a sequence of methods. The existing methods of road semantic segmentation methods are designed from two different aspects: high-accuracy models (*e.g.*, LinkNet [20], PSPNet [21], ResUNet [22], DeepLabv3+ [23], MSAD-Net [24]) and light-weight models (*e.g.*, ContextNet [25], FastSCNN [26], LRSR-net [27], and CGNet [28]). Among high-accuracy models, U-Net [29] and LinkNet are the highly regarded encoder-decoder structures for road semantic segmentation. Specifically, both LinkNet and U-Net share similar network structures that consist of a sequence of down-sampling layers (*Encoder*), a series of up-sampling layers (*Decoder*), and lateral connections. The key difference lies in LinkNet optimizes the utilization of network parameters. Thereafter, Zhang et al. integrated residual learning and U-Net to design a road semantic segmentation neural networks called ResUNet [22] and achieved good results. A highly esteemed network framework, DeepLabv3+, proposed a combination of spatial pyramid polling module and encoder-decoder structure to construct a state-of-the-art road semantic segmentation model that has been unrivaled for an extended period. Although these methods exhibit high-accuracy, they have a large number of parameters such as DeepLabv3+ of 59.33M and PSPNet of 65.57M, so that they are not directly transferable to real-time applications or embedded devices.

To avoid the above dilemma, many light-weight road extraction models have been proposed. For instance, Poudel et al. reported a new deep neural network framework, ContextNet, which builds on factorized convolution, network compression and pyramid representation to real-time detect road with low computational cost [25]; Sun et al. presented a dilated joint convolution module to enhance the extraction of local features while reducing the overall model parameters [27]. Their model's accuracy rate increased by at least 3% compared with the existing advanced detection methods; and in CGNet, Wu et al. proposed the context guided block to learn the joint feature of both local feature and surrounding context, and improve the joint feature with the global context [28]. Unfortunately, these methods with a small footprint suffer from low accuracy in road extraction due to their adherence to image classification design principles while ignoring the inherent properties of road semantic segmentation.

On the other hand, several technologies are proposed to enhance road detection or traffic accident detection [30], [31], [32] by using a combination of 2D and 3D information. For example, a pavement crack detection method that integrates 2D grayscale images and 3D laser scanning data based on Dempster-Shafer theory is proposed [33]. Hu et al. [34] employed both 2D and 3D information from the images taken at high speed to detect 3D pavement defects. Bayoudh et al. [35] introduced hybrid 2D-3D CNNs models based on the transfer learning for traffic sign recognition and semantic road detection. Nevertheless, this work merely combines 2D and 3D CNNs and consider little about the correlations between hierarchical features from 2D and 3D CNNs. Furthermore, the aforementioned techniques demonstrate high effectiveness in normal lighting conditions, however, their efficacy will be significantly diminished in low light environments, particularly during nighttime.

Unlike the intensity information used in conventional vision systems, the property of polarization offers intriguing physical characteristics of light, enabling the extraction of distinctive information about an object, such as its surface smoothness, 3D normal, and material composition. To push the envelope further, we exploit the design ethos of semantic segmentation and propose a novel model architecture by analyzing comprehensively multiple polarization characteristics, 2D plane, and 3D space, which is a light-weight network specially tailored for full-time road detection to achieve high accuracy.

C. Basic Knowledge of Polarization

High-dimensional optical information, especially polarization as an electromagnetic physical characteristic, is vital for comprehensive non-invasive characterization of targets in different scenarios [36], [37], [38]. Techniques that image the polarization, also known as polarization parameters, have aroused wide applications in the domains of remote sensing, marine resources detection, *etc* [39], [40], [41], [42], [43], [44]. Optical polarization information often is characterized by Stokes vector S [45], which can be obtained using polarization imaging system, *e.g.*, LWIR DoFP polarization camera whose surface integrates a polarization modulator consisting of micro-polarizers with four different polarization orientations of 0°, 45°, 90°, and 135°. The Stokes parameters can be expressed as:

$$S = \begin{bmatrix} S_0 \\ S_1 \\ S_2 \end{bmatrix} = \begin{bmatrix} I_{0^\circ} + I_{90^\circ} \\ I_{0^\circ} - I_{90^\circ} \\ I_{45^\circ} - I_{135^\circ} \end{bmatrix} \tag{1}$$

where S_0 refers to the total intensity received by the camera; S_1 represents the intensity difference between the vertical and horizontal components; S_2 denotes the intensity difference between the 45° and 135° components. According to Eq. (1), the degree of polarization (DoP) and the angle of polarization

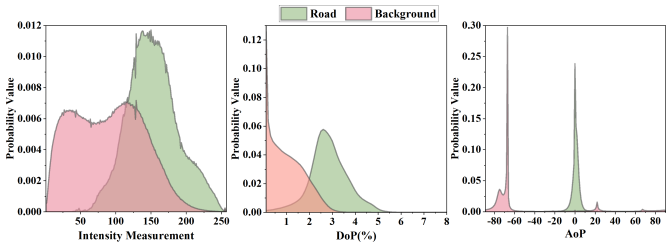

Fig. 2. The probability distributions of S_0, AoP, and DoP of the road and the background for all images in the LDDRS dataset.

(AoP) can be described as:

$$\mathrm{DoP} = \frac{\sqrt{{S_1}^2 + {S_2}^2}}{S_0} \tag{2}$$

$$\mathrm{AoP} = \frac{1}{2}\arctan\left[\frac{S_2}{S_1}\right] \tag{3}$$

Previous studies [9], [10], [11], and [12] have provided a detailed analysis of the road polarization characteristics in LWIR. Herein three key points are summarized: **1**) The polarization state of a road is dominated by the polarized emitted light, and the AoP of the road is close to zero with respect to the horizontal plane; **2**) As the physical properties such as surface smoothness, radiation condition, and normal direction of the road regions are similar, the AoP is also similar; **3**) Roads and backgrounds have different discriminative polarization information. Fig. 2 also presents the probability distributions of the S_0, DoP, and AoP for both road and background in 2113 images from the LDDRS dataset [11]. By examining Fig. 2, it is evident that the majority of AoP values for the road are in close proximity to zero, and there exist significant disparities between the probability distribution curves for both the road surface and background. This indicates that DoP and AoP are more suitable for road detection.

III. Lightweight Multi-Pathway Collaborative 2D/3D Convolutional Networks for Full-Time Road

As shown in Fig. 3, we develop an LMPC2D3DCNet, which is an infrared polarization-empowered lightweight networks for full-time road detection. In this section, we will elaborate the LMPC2D3DCNet architecture.

A. Overview of Our Approach

The overall pipeline of our proposed LMPC2D3DCNet is illustrated in Fig. 3, which mainly consists of five components. Starting from the input LWIR polarization images (LWIR-PIs) of nighttime or daytime, LMPC2D3DCNet first employs a BM3D and a polarization difference model to denoise and de-mosaic respectively. Meanwhile, features extracted from LWIR-PIs, S_0, DoP, and AoP are fed into Shallow Feature Extraction (SFE) module and subsequently input into MPC2D3DCNet to distill the relationship between polarization characteristics and space. The SFE module includes two collaborative 2D/3D CNN layers (refer to Section III-C for more details) with a kernel size of 3×3. Finally, MPC2D3DCNet performs the road detection using Road Detection module that comprises two collaborative 2D/3D CNN layers with a kernel size of 1×1. Our designed enhanced loss function module, which is composed of Content Loss, Adversarial Loss, and Texture Gradient Loss, is employed to study the model. Note that the Gradient module and Pre-Processing stage will be described in the Section III-E and IV-A, respectively.

B. Multi-Pathway Collaborative 2D/3D Convolutional Networks

LMPC2D3DCNet is designed to automatically identify road region through distilling the relationship between polarization characteristics and 2D plane or 3D space from multi-pathway collaborative 2D/3D convolutional networks (MPC2D3DCNet). We show the detailed network structure of MPC2D3DCNet in Figs. 4 and 5. The proposed MPC2D3DCNet mainly contains three parts: multi-pathway encoder, multi-pathway decoder, and cross-space skip connection.

1) Multi-Pathway Encoder (MPE): Inspired by the collaboration of diverse neural cells in achieving a vital function, the *MPE*, whose structure is shown in Fig. 4, is constructed by five types of neural cells, as depicted in Fig. 5. The modules in Figs. 5 (a) and (b) have a single 2D or 3D block, while the modules in Figs. 5 (c), (d) and (e) contain both the 2D and 3D blocks. Moreover, the module in Fig. 5 (e) is also equipped with a new cross 2D-3D non-local attention network (C2D3DNLA). The 2D or 3D block consists of two 2D or 3D convolutional layers, followed by the instance normalization and the LeakyReLU activation function. The MaxPooling operation is employed during the encoding stage to reduce spatial resolution and aggregate long-range information. Specifically, *MPE*'s structure is a standard triangle, which consists of two stems and cross-space information transfer neural cells, *i.e.*, the module in Fig. 5(e). Both stems primarily transmit 2D and 3D information that corresponds to the right and left data stream pipes of *MPE*. The interaction between 2D and 3D information of the branches on the two stems through a collaborative 2D/3D CNN block, which then propagates to more distant branches. Its primary goal is to distill more discriminative features.

2) Multi-Pathway Decoder (MPD): Symmetrically, the *MPD*, whose structure adopted in [46] is shown in Fig. 6, consists of three types of neural cells. Fig. 7 shows the basic modules of the *MPD*. The modules in Figs. 7 (a) and (b) have a single 2D or 3D block, while the module in Fig. 7 (c) includes both the 2D and 3D blocks. The *MPD* adopts linear interpolation to gradually recover the spatial resolution of feature maps.

3) Cross-Space Skip Connection: How to harness *MPE* and *MPD* through skip connection is of great significance in relation to stimulating the strength of model effectively. We incorporate skip connections (*e.g.*, $MPE^a \dashrightarrow MPD^a$) to facilitate the exchange of low-level information between the encoder and decoder, thereby preserving low-level features and achieving multiscale feature fusion. Herein the overall process

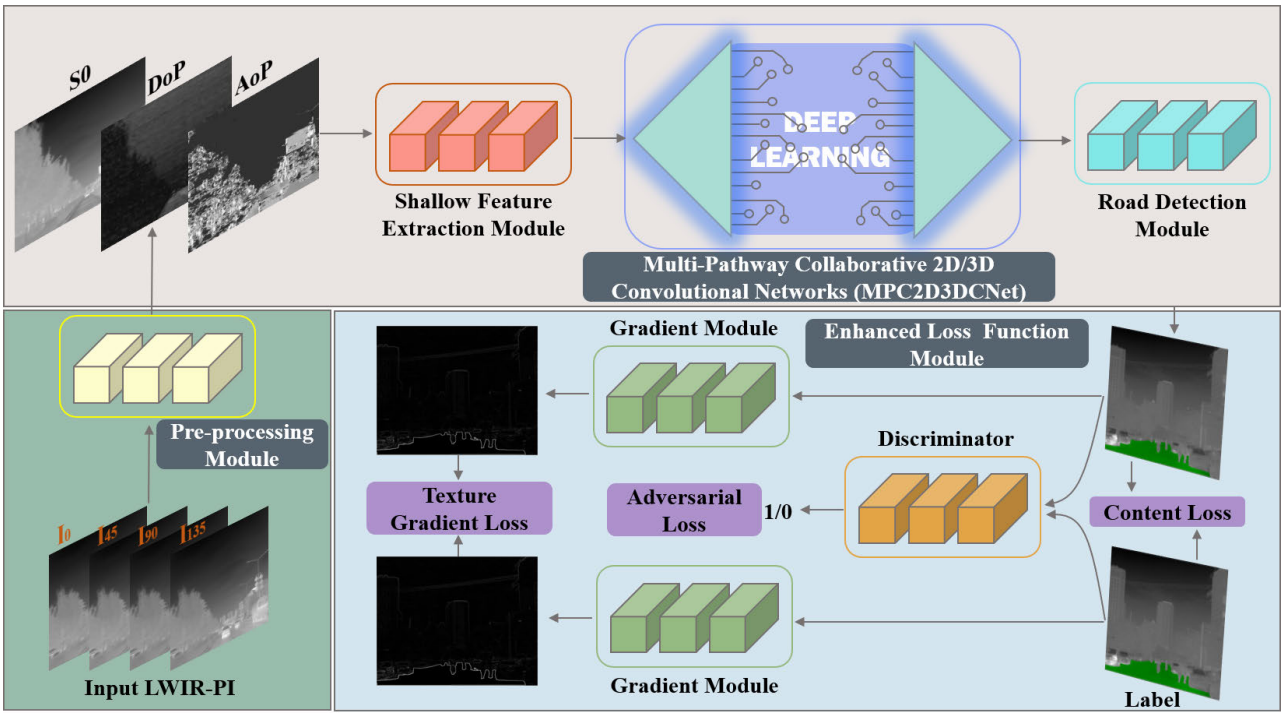

Fig. 3. Overall architecture of the proposed LMPC2D3DCNet. LMPC2D3DCNet is composed of three main parts: the pre-processing module, the multi-pathway collaborative 2D/3D convolutional networks (MPC2D3DCNet), and the enhance loss function module.

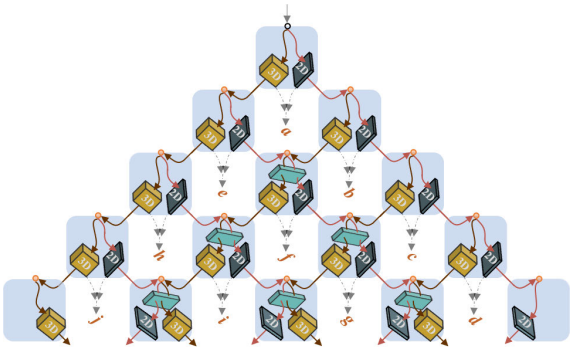

Fig. 4. The architecture of the multi-pathway encoder.

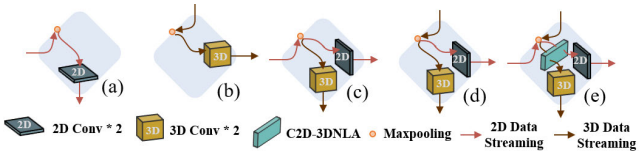

Fig. 5. The basic modules of the multi-pathway encoder.

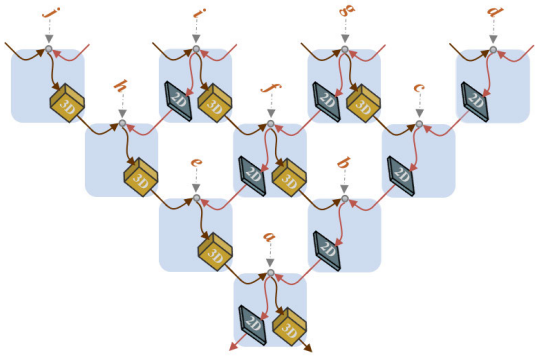

Fig. 6. The architecture of the multi-pathway decoder.

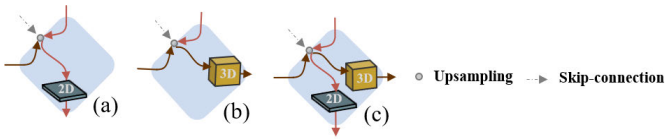

Fig. 7. The basic modules of the multi-pathway decoder.

of skip connections can be expressed as:

$$\boldsymbol{F} = Concat\left[\boldsymbol{F}_{skip}^{2D} \ominus \boldsymbol{F}_{skip}^{3D}, \boldsymbol{\mathcal{U}}\left(\boldsymbol{F}^{\mathbf{2D}}\right) \ominus \boldsymbol{\mathcal{U}}\left(\boldsymbol{F}^{\mathbf{3D}}\right)\right] \tag{4}$$

where $\boldsymbol{\mathcal{U}}(\cdot)$ denotes a trilinear interpolation upsampling operation, and $\ominus$ represents element-wise subtraction.

Remarkably, the designed MPC2D3DCNet has three advantages: i) MPC2D3DCNet integrates features from multiple pathways, enabling the learning of multi-level representations and enhancing expressive capacity; ii) MPC2D3DCNet incorporates numerous cross-space skip connections that facilitate the propagation of information, particularly low-frequency information and gradients; iii) MPC2D3DCNet exploits the merit of 2D (high detection accuracy of the easily recognized regions in 2D view) and 3D (high smoothness of 3D scene contour) representations simultaneously to model road regions [46].

C. Collaborative 2D/3D CNN

The main idea behind collaborative 2D/3D CNN module (C2D3DC) is to balance the semantic extraction gap

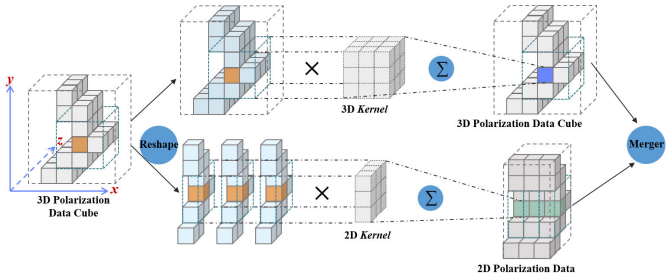

Fig. 8. Collaborative 2D/3D convolutional neural network.

between sparse inter-channel polarization information and dense intra-channel polarization feature information. The 2D CNNs only capture features in the *x-y* plane with even receptive fields (*e.g.*, the spatial information of DoP), resulting in severe underfitting of inter-slice features. On the other hand, although 3D CNNs can capture inter-slice correlations, they suffer from overrepresentation for highly discontinuous long-range sparse information along the $z-$axis, which makes the network prone to over-fitting. We propose an elegant solution to remedy this issue by combining 2D and 3D convolutional neural networks. Fig. 8 illustrates detailed collaborative architecture. Specifically, given a group 3D feature $P \in \mathbb{R}^{C\times D\times H\times W}$ composed of S_0, DoP, and AoP, the overall process of C2D3DC is summarized as:

$$\bar{P}_{2D} = f_{\text{conv2D}}\ (\mathcal{R}\ (P)) \tag{5}$$

$$\bar{P}_{3D} = f_{\text{conv3D}}\ (P) \tag{6}$$

$$\bar{P}_O = abs\ (\mathcal{R}\ (\bar{P}_{2D}) \cup \bar{P}_{3D} - \mathcal{R}\ (\bar{P}_{2D}) \cap \bar{P}_{3D}) \tag{7}$$

where $f_{\text{conv2 D}}(\cdot)$ and $f_{\text{conv3 D}}(\cdot)$ denote 2D and 3D convolution operations, respectively; $\mathcal{R}(\cdot)$ represents *Reshape* operation. Eq. (7) is utilized to discard the redundant features and decrease computational overhead.

D. Cross 2D-3D Non-Local Attention

As a matter of fact, convolutional operation only processes local neighborhood information. In this work, inspired by non-local neural networks [47], we proposed cross 2D-3D non-local attention (C2D3DNLA) for capturing long-range dependencies between 2D plane and 3D space. A schematic of C2D3DNLA is shown in Fig. 9.

The C2D3DNLA consists of two parts: channel non-local attention (CNLA) and spatial non-local attention (SNLA). Given two input feature maps $\boldsymbol{F}^{2D} \in \mathbb{R}^{C\times H\times W}$ and $\boldsymbol{F}^{3D} \in \mathbb{R}^{C\times D\times H\times W}$, the output of the CNLA ($\boldsymbol{F}$') and the output of SNLA ($\boldsymbol{F}$") can be expressed as:

$$\boldsymbol{F}' = M_c\ (\boldsymbol{F^{2D}};\boldsymbol{F^{3D}}) \oplus \boldsymbol{F^{3D}} \tag{8}$$

$$\boldsymbol{F}'' = M_s\ (\boldsymbol{F^{2D}};\boldsymbol{F^{3D}}) \oplus \boldsymbol{F^{3D}} \tag{9}$$

where $M_c\ (\boldsymbol{F^{2D}};\boldsymbol{F^{3D}})$ represents CNLA map, $M_s\ (\boldsymbol{F^{2D}};\boldsymbol{F^{3D}})$ represents SNLA map, and $\oplus$ denotes element-wise sum. $M_c\ (\boldsymbol{F^{2D}};\boldsymbol{F^{3D}})$ and $M_s\ (\boldsymbol{F^{2D}};\boldsymbol{F^{3D}})$ can be calculated as follows:

$$M_c\ (\boldsymbol{F^{2D}};\boldsymbol{F^{3D}}) = \mathcal{S}(\mathcal{T}(\boldsymbol{F}_R^{2D}) \otimes \boldsymbol{F}_R^{3D}) \otimes \boldsymbol{F}_R^{3D} \tag{10}$$

$$M_s(\boldsymbol{F}^{2D};\ \boldsymbol{F}^{3D}) = \mathcal{S}(\mathcal{A}(\boldsymbol{F}_R^{2D}) \otimes \mathcal{T}(\boldsymbol{F}_R^{3D})) \otimes \mathcal{A}(\mathcal{T}(\boldsymbol{F}_R^{3D})) \tag{11}$$

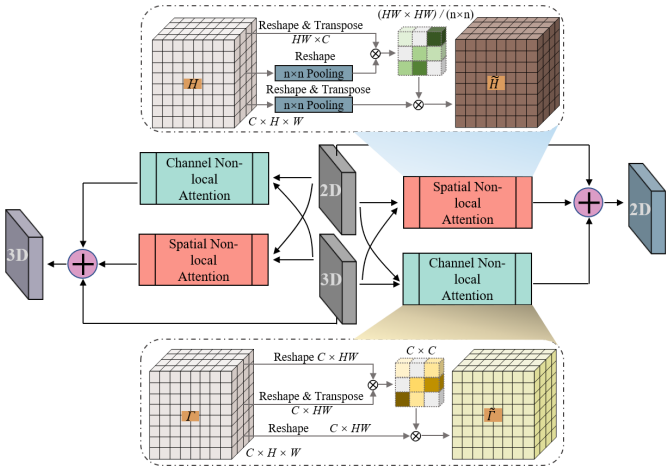

Fig. 9. The cross 2D-3D non-local attention module.

where $\boldsymbol{F}_R^{2D}$ *and* $\boldsymbol{F}_R^{3D}$ represent the $\boldsymbol{F^{2D}}$ and $\boldsymbol{F^{3D}}$ operated by the *Reshape* operation respectively; $\mathcal{S}(\cdot)$, $\mathcal{T}(\cdot)$, and $\mathcal{A}(\cdot)$ represent *Softmax*, *Transpose*, and *AvgPooling* operations; $\otimes$ denotes the multiplication of corresponding elements.

E. Enhanced Loss Function

Unlike existing methods, we adopt three losses to guide objective function $\mathcal{L}_{obj}$: a content loss $\mathcal{L}_{con}$, a texture gradient loss $\mathcal{L}_{teg}$, and an adversarial loss $\mathcal{L}_{adv}$:

$$\mathcal{L}_{obj} = \lambda_{con} \cdot \mathcal{L}_{con} + \beta_{teg} \cdot \mathcal{L}_{teg} + \mu_{adv} \cdot \mathcal{L}_{adv} \tag{12}$$

where λ_{con}, β_{teg}, and μ_{adv} are employed to adjust the wights of different loss items. Increasing λ_{con}, β_{teg}, and μ_{adv} can get better perceived quality road regions, and vice versa.

1) Content Loss $\mathcal{L}_{con}$: *Content Loss* $\mathcal{L}_{co}$ is designed to improve the objective quality of the identified road by minimizing the L_2-norm of error. In this work, it includes two pixel-level reconstruction losses, *i.e.*, L_2 loss $\mathcal{L}_2$ and perceptual loss $\mathcal{L}_{per}$ [48]:

$$\mathcal{L}_2 = \left\| I_{gt} - I_{det} \right\|_2^2 \tag{13}$$

$$\mathcal{L}_{per} = \left\| \Phi(I_{gt}) - \Phi(I_{det}) \right\|_2^2 \tag{14}$$

$$\mathcal{L}_{con} = \mathcal{L}_2 + \mathcal{L}_{per} \tag{15}$$

where I_{gt} and I_{det} mean the ground-truth and detected road images; $\|\cdot\|_2$ denotes the L_2-norm; Φ denotes the feature extractor of VGG19 [49].

2) Texture Gradient Loss $\mathcal{L}_{teg}$: The *Content Loss* $\mathcal{L}_{co}$ can makes the detected road regions close to the ground-truth in pixel-level, thus mitigating the non-smoothness problem. Nevertheless, it neglects the local variation relationship of the pixel values, leading to an overly smooth generated road region. To bring out more clearer edges in the detected road images, preserving the local variation relationship of image pixels is crucial. The gradient of an image can describe the local change of pixel value, usually expressed as the edge and texture details. Fig. 10 shows this problem, where the gradient image is calculated based on gradient module. The gradient image has more detailed information, and thereby the high-quality

Fig. 10. Comparison of LWIR S_0 image and gradient image.

road edge can be obtained. Therefore, we incorporate a texture gradient loss term into the objective function.

$\mathcal{L}_{teg}$ works on the gradient images to learn road regions with better perception quality and object quality, which can be expressed as

$$\mathcal{L}_{teg} = \left\| S_{obel}(I_{gt}) - S_{obel}(I_{det}) \right\|_2^2 \tag{16}$$

$$S_{obel}(\Theta) = [\nabla_x, \nabla_y]^T \tag{17}$$

where $\nabla_x = Conv^{2D}(x, \mathcal{K})$ and $\nabla_y = Conv^{2D}(y, \mathcal{K}^T)$ denote the gradient information in the horizontal direction and vertical direction, respectively; $S_{obel}(\cdot)$ is *Sobel* operator. $\mathcal{K}$ is the parameters of the 2D convolution kernel. In this work, we set $\mathcal{K}$ as

$$\mathcal{K} = \begin{bmatrix} 1 & 0 & -1 \\ 2 & 0 & -2 \\ 1 & 0 & -1 \end{bmatrix} \tag{18}$$

3) Adversarial Loss $\mathcal{L}_{adv}$: The solution is highly ill-posed as the detection result derived by minimizing L_2-norm is equivalent to average value of multiple potentially practical feasible solutions. Thus, using $\mathcal{L}_{con}$ and $\mathcal{L}_{teg}$ solely cannot ensure texture detail information rich enough. To remedy this issue, standard GAN loss function is introduced for minimizing the loss between reconstructed and ground-truth images, and written as,

$$\mathcal{L}_{adv} = \mathbb{E}gt\left[\log D(I_{gt})\right] + \mathbb{E}rec[\log(1 - D(I_{det}))] \tag{19}$$

where structure of discriminator D is a PatchGAN [50]; $E[\cdot]$ represents the expectation. It further improves the road detection and meanwhile maintains the smooth edges quality.

IV. Result and Discussion

In this section, we will first provide a description of the dataset and then present the training details of LMPC2D3DCNet. Next, we will introduce the evaluation metrics. Subsequently, we will perform both qualitative and quantitative assessments of the results produced by our LMPC2D3DCNet and compare it with the state-of-the-art methods. Finally, series of ablation studies are conducted to demonstrate the effectiveness of each component in LMPC2D3DCNet. Additionally, we will also analyze the efficiency and failure cases.

A. Dataset

In order to experimentally demonstrate the effectiveness of our proposed LMPC2D3DCNet, we train and verify it on a LWIR DoFP dataset of road scene (LDDRS) [11], which is currently the only publicly available high-quality dataset. The images in LDDRS are photographed using an uncooled infrared DoFP camera with 512×640 resolution in 14 bits. The dataset consists of 2113 images, which provides both infrared intensity (S_0) and polarization information. In the pre-processing stage, all images are first denoised using the BM3D [51] and de-mosaiced using a polarization difference model [52]. We then calculate S_0, DoP, and AoP as described in Eqs. (1), (2) and (3). The road regions of all images are manually annotated, covering urban road and highway both daytime and nighttime. Also, the LDDRS includes different traffic situations, *e.g.*, different numbers of cars and pedestrians in the road scenario, and many shadow regions and areas obscured by trees, which can effectively validate our proposed method. For more detailed information on the dataset construction, please refer to ref [11]. Finally, we randomly selected 1960 images and 106 images as the training set and validation set (VAL106), respectively. The remaining 317 images constitute the testing set (TEST317).

B. Training Setting

All experiments are carried out on Linux Server Intel Core i7-7700 CPU @3.6Hz 48.0GB of RAM, and Python 3.7 programming. The proposed LMPC2D3DCNet are implemented in the PyTorch framework [53]. The trade-off parameters in Eq. (12) λ_{con}, β_{teg}, and μ_{adv} are empirically set as 1, 0.10, and 0.005, respectively, with numerous experiments. At the training procedure, we adopt the Adam algorithm as the optimizer to optimize the model parameters with momentum term ($\beta_1 = 0.9$ and $\beta_2 = 0.999$). Note that we do not employ any data augmentation techniques. The initial learning rate is set to 0.001, and the learning rate is reduced to half of the original every 15 epochs. The models are trained on a single Nvidia GeForce RTX 3090 GPU with a min-batch size of 8.

C. Evaluation Metrics

Seven widely recognized metrics are employed to comprehensively evaluate the performance of our method for road detection, *i.e.*, precision (*PRE*), recall (*REC*), overall accuracy (*OA*), Matthew's correlation coefficient (*MCC*), $F1-$Score (*F1*), Intersection over Union (*IoU*), and balanced error rate (*BER*). Among them, the larger the values of *PRE*, *REC*, *OA*, *MCC*, *F1*, and *IoU*, the better the road surface detection performance, and the smaller the *BER*, the better the detection results. The *MCC* measures the correlation between the expected class and the predicted class. These evaluation indexes are calculated as shown in the following equations:

$$PRE = \frac{TP}{TP + FP} \tag{20}$$

$$REC = \frac{TP}{TP + FN} \tag{21}$$

$$OA = \frac{TP + TN}{TP + FP + TN + FN} \tag{22}$$

$$MCC = \frac{TP \times TN - FP \times \mathrm{FN}}{\sqrt{(TP + FP) \times (\mathrm{TP} + \mathrm{FN}) \times (\mathrm{TN} + \mathrm{FP}) \times (\mathrm{TN} + \mathrm{FN})}} \tag{23}$$

TABLE I

QUANTITATIVE EVALUATIONS WITH THE STATE-OF-THE-ART METHODS ON LDDRS TESTING DATASET, NUMBER HIGHLIGHTED WITH RED, GREEN, AND BLUE TO INDICATE THE BEST THREE RESULTS. ↑&↓ DENOTE LARGER AND SMALLER IS BETTER, RESPECTIVELY

Method	*PRE*(%) ↑	*REC*(%) ↑	*OA*(%) ↑	*MCC*(%)↑	*IoU*(%) ↑	*BER*(%) ↓	*F1*(%) ↑	*P-value* $^{(IoU)}$	*P-value* $^{(F1)}$
SegNet	98.61	76.09	98.04	85.26	75.32	12.00	85.14	<0.0001**	<0.0001**
LinkNet	98.85	85.95	98.62	91.29	85.28	7.06	91.59	<0.0001**	<0.0001**
UNet	95.07	91.60	98.88	92.40	87.85	4.40	92.67	<0.0001**	<0.0001**
ResUNet	97.07	93.74	99.23	94.88	91.18	3.27	95.20	<0.0001**	<0.0001**
Unet++	96.45	93.06	99.08	93.94	90.01	3.64	94.18	<0.0001**	<0.0001**
UNet3+	95.73	93.07	99.01	93.73	89.76	3.68	94.10	<0.0001**	<0.0001**
CGNet	94.92	94.89	99.10	94.32	90.36	2.82	94.73	<0.0001**	<0.0001**
DABNet	96.27	95.15	99.25	95.27	91.79	2.62	95.64	<0.0001**	<0.0001**
LRSR-net	97.49	94.14	99.30	95.36	92.11	3.05	95.66	0.0002*	0.0002*
ContextNet	95.60	95.00	99.17	94.74	91.06	2.73	95.07	<0.0001**	<0.0001**
DeepLabv3 ~	98.70	89.81	98.97	93.54	88.86	5.14	93.91	<0.0001**	<0.0001**
DeepLabv3 #	97.43	91.89	99.07	94.01	89.80	4.18	94.36	<0.0001**	<0.0001**
ENet	95.22	94.95	99.12	94.48	90.61	2.76	94.82	<0.0001**	<0.0001**
PSPNet	99.64	91.85	98.95	93.38	88.73	4.26	93.82	<0.0001**	<0.0001**
DenseASPP	90.72	92.41	98.68	90.67	84.45	4.21	91.19	<0.0001**	<0.0001**
MSADNet	89.56	94.74	98.72	91.25	85.42	3.11	91.71	<0.0001**	<0.0001**
FastSCNN	96.51	95.08	99.24	95.30	91.94	2.65	95.63	<0.0001**	<0.0001**
ERFNet	97.32	71.80	96.95	82.03	70.36	14.19	82.46	<0.0001**	<0.0001**
LaneNet	98.43	92.37	99.20	94.89	91.05	3.89	95.24	<0.0001**	<0.0001**
LDNet	97.43	94.34	99.29	95.39	92.06	2.95	95.65	<0.0001**	<0.0001**
DDRNet	94.84	95.60	99.15	94.72	90.92	2.46	95.15	<0.0001**	<0.0001**
Ours	96.96	96.68	99.45	96.46	93.85	1.80	96.72	-	-

~ DeepLabv3 means DeepLabv3+(Xception); # DeepLabv3 denotes DeepLabv3+(ResNet101); *P-value* $^{(IoU)}$ and *P-value* $^{(F1)}$ denote that significance (*p*-value) in the *IoU* and *F1* values between our LMPC2D3DCNet and other methods are assessed by Wilcoxon rank-sum test, respectively. *P-value<0.001, **P-value<0.0001.

$$F1 = \frac{2TP}{2TP + FP + FN} \tag{24}$$

$$IoU = \frac{TP}{TP + FP + FN} \tag{25}$$

$$BER = 1 - \frac{1}{2} \times \left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP}\right) \tag{26}$$

where *TP* (true positive) refers to the number of correctly identified road regions at the pixel-level; *FP* (false positive) means the number of not correctly identified road regions at the pixel-level; *TN* (true negative) means the number of correctly identified background regions at the pixel-level; *FN* (true negative) means the number of not correctly identified background regions at the pixel-level.

In addition, we adopt Parameters (*Params*) and floating point operations (*FLOPs*) to evaluate the complexity of different methods.

D. Comparison With State-of-the-Art Methods

In order to demonstrate the effectiveness of the proposed LMPC2D3DCNet, we conduct comparative experiments on the LDDRS dataset. The road detection results of LMPC2D3DCNet are compared with the other state-of-the-art methods, including SegNet [15], LinkNet [20], UNet [29], ERFNet [54], ResUNet [22], Unet++ [55], UNet3+ [56], DABNet [17], LRSR-net [27], ContextNet [25], DeepLabv3+(Xception) [23], DeepLabv3+(ResNet101) [23], ENet [13], PSPNet [21], DenseASPP [14], MSADNet [24], FastSCNN [26], LaneNet [57], LDNet [58], and DDRNet [59]. For an objective and fair comparison, all methods are trained from scratch on the training set of LDDRS. During the test phase, the testing set containing daytime and nighttime is used and nine metrics mentioned above are counted.

1) Quantitative Evaluation: The quantitative experimental results of LMPC2D3DCNet and other methods are presented in Table I. It is straightforward to find from Table I that LMPC2D3DCNet consistently surpasses other state-of-the-art methods concerning all seven evaluation indexes except for *PRE*. More specifically, compared with the second-best method LRSR-net, the *REC*, *OA*, *MCC*, *IoU*, *BER*, and *F1* values of LMPC2D3DCNet are 96.68%, 99.45%, 96.46%, 93.85%, 1.80%, and 96.72%, respectively, which are improvement of approximately 2.70%, 0.15%, 1.15%, 1.89%, 40.98%, and 1.11% over LRSR-net, respectively. It has not escaped

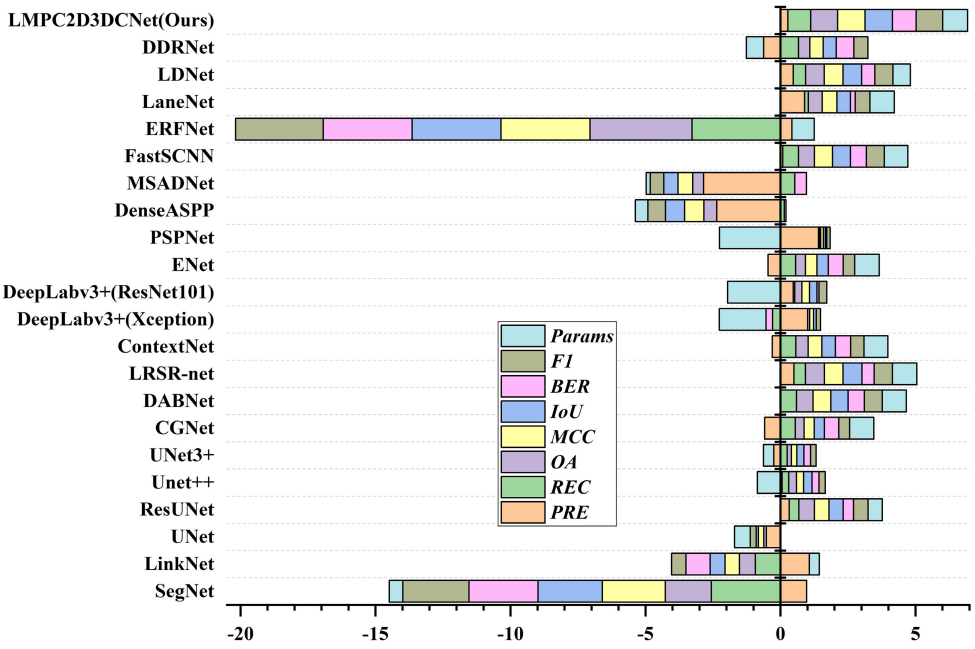

Fig. 11. Ranking of the methods in the global performance evaluation. Our proposed LMPC2D3DCNet and other methods are ranked according to the sum of the Z-scores of all the evaluation metrics on the LDDRS test dataset.

from our notice that although PSPNet obtains the highest *PRE* (99.64%), its *OA*, *MCC*, *IoU*, *BER*, and *F1* scores, which can better reflect the overall performance, are 0.51%, 3.30%, 5.77%, 57.75% and 3.09% lower than those of LMPC2D3DCNet, respectively. It is noteworthy that the proposed LMPC2D3DCNet is the sole approach among 22 methods that achieves $OA>99.45$, $MCC>96.46$, $IoU>93.85$, $BER<1.80$, and $F1>96.72$ simultaneously on the testing set of LDDRS. Furthermore, Wilcoxon's rank sum test [60] is used to test for differences in the distributions of *IoU* and *F1* values between our LMPC2D3DCNet and other compared methods. Specifically, the differences between our method and other methods in the *IoU* and *F1* values are statistically significant, which have a $p-$value <0.001 in the Wilcoxon's rank sum test.

Since the source codes of PCRL [9] and PolarNet [12] are not publicly available, we have extracted their results from their respective literature. Concretely, the *PRE*, *REC*, and *IoU* of LMPC2D3DCNet are 3.80%, 1.73%, and 5.31% higher than those reported by PCRL, respectively; 1.33%, 0.46%, and 1.71% higher than those reported by PolarNet, respectively.

We also rank these methods using the sum of Z-scores of all evaluation indexes to analyze the comprehensive performance of various road detection methods. As shown in Fig. 11, it can be found that LMPC2D3DCNet yields the best comprehensive performance among all methods.

2) Qualitative Comparison: The visual comparison between LMPC2D3DCNet and the other comparison methods is shown in Fig. 12. Here we select six methods for illustration. The first five columns and the last five columns show a comparison of different methods of road detection during daytime and nighttime, respectively. By carefully observing Fig. 12, the following four phenomena can be seen: i) When the shadow on the bottom of the car obscures the road (see first column), the proposed LMPC2D3DCNet can achieve better road detection results that extremely close to the ground truth. However, the other methods all suffer from serious false detection in this case. ii) The second and tenth columns reveal that other methods tend to misclassify car glass as road, while LMPC2D3DCNet exhibits relatively superior performance. iii) The comparison results at the edges of road or car detection are presented in columns three, seven, eight, and nine. These cases show that LMPC2D3DCNet enjoys smoother results and exhibits superior performance at the edge regions of the road. Conversely, ResUNet performs most poorly among these methods. iv) LMPC2D3DCNet also exhibits a stable generalization capability across daytime and nighttime.

E. Model Analysis

This subsection provides the advantages of our proposed MPC2D3DCNet by quantitative analysis.

1) Effect of Different Loss Configurations: To research the influences of various loss configurations, we train the proposed LMPC2D3DCNet with different losses (*i.e.*, Binary Cross-entropy (BCE), MSE loss $\mathcal{L}_2$, perceptual loss $\mathcal{L}_{per}$, *Texture Gradient Loss* $\mathcal{L}_{teg}$, standard adversarial loss $\mathcal{L}_{adv}$, and their combinations). We employ the BCE as the comparison baseline. Table II summarizes the performance comparison of different losses on both LDDRS validation and testing datasets. As shown in Table II, the performance of $\mathcal{L}_2+\mathcal{L}_{per}+\mathcal{L}_{teg}+\mathcal{L}_{adv}$ is superior to those of the other losses. Taking results of testing set as an example, $\mathcal{L}_2+\mathcal{L}_{per}+\mathcal{L}_{teg}+\mathcal{L}_{adv}$ reaches 0.88%, 9.15%, 60.66%, and 5.61% average enhancements of *OA*, *IoU*, *BER*, and *F1* scores, respectively, compared to the other four loss configurations. The results demonstrate that $\mathcal{L}_2+\mathcal{L}_{per}+\mathcal{L}_{teg}+\mathcal{L}_{adv}$ is much more suitable for road detection task.

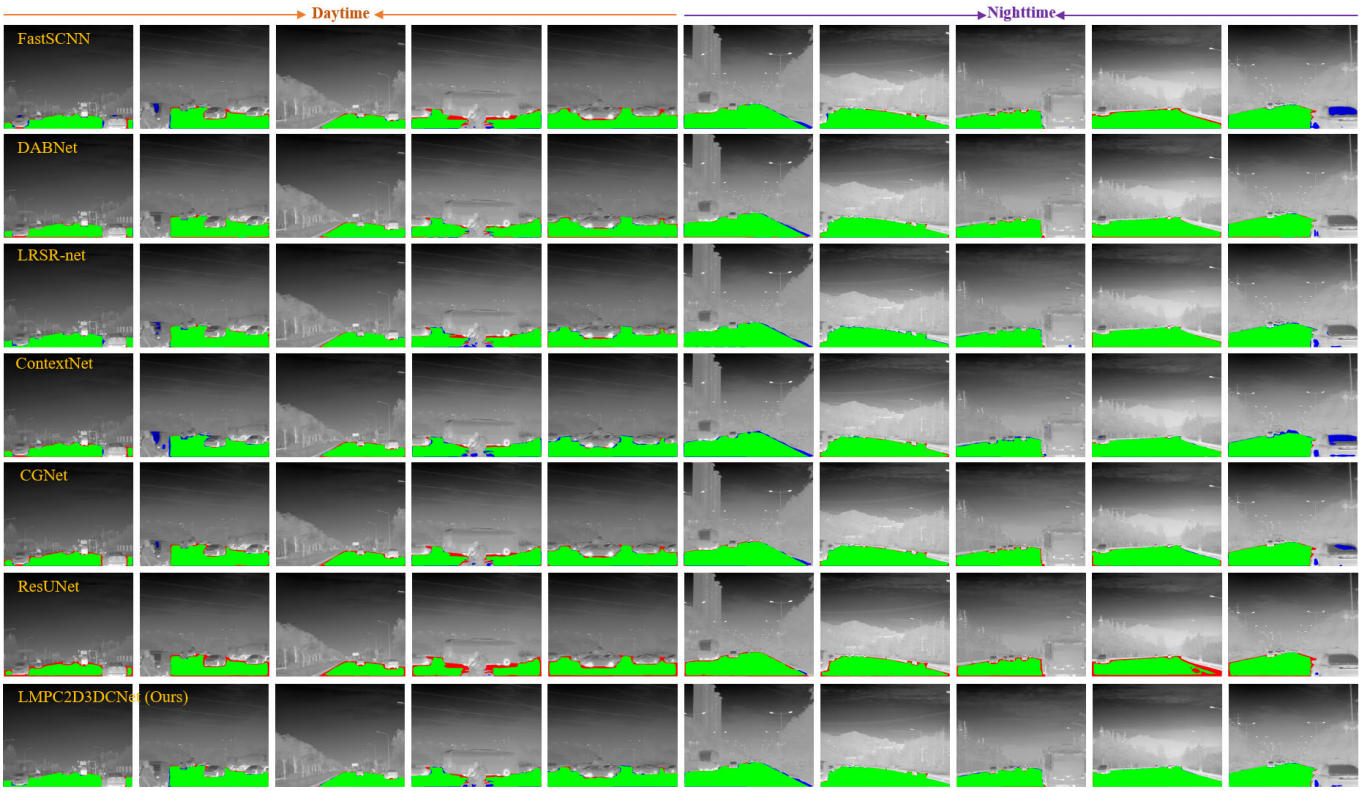

Fig. 12. Detection results of different comparison methods. Green regions, blue regions, and red regions represent true positive, false positive, and false negative, respectively.

TABLE II

PERFORMANCE COMPARISON OF DIFFERENT LOSS CONFIGURATIONS ON BOTH LDDRS VALIDATION DATASET AND TESTING DATASET, NUMBER HIGHLIGHTED WITH RED, GREEN, AND BLUE TO INDICATE THE BEST THREE RESULTS. ↑&↓ DENOTE LARGER AND SMALLER IS BETTER, RESPECTIVELY

Dataset	Loss Function	*PRE*(%) ↑	*REC*(%) ↑	*OA*(%) ↑	*MCC*(%) ↑	*IoU*(%) ↑	*BER*(%) ↓	*F1*(%) ↑
VAL106	BCE	99.74	78.46	98.56	87.67	78.29	10.77	87.61
	$\mathcal{L}_2$	94.26	89.98	99.13	91.54	85.44	5.15	91.85
	$\mathcal{L}_2$+$\mathcal{L}_{per}$	96.13	90.10	99.23	92.56	87.00	5.06	92.78
	$\mathcal{L}_2$+$\mathcal{L}_{per}$+$\mathcal{L}_{teg}$	92.86	94.31	99.32	93.14	88.11	3.04	93.40
	$\mathcal{L}_2$+$\mathcal{L}_{per}$+$\mathcal{L}_{teg}$+$\mathcal{L}_{adv}$	98.10	93.22	99.28	95.17	91.60	3.47	95.44
TEST317	BCE	99.54	76.32	97.02	83.88	75.97	11.85	83.74
	$\mathcal{L}_2$	99.10	87.83	98.94	92.61	87.12	6.12	92.86
	$\mathcal{L}_2$+$\mathcal{L}_{per}$	96.97	92.96	99.09	94.37	90.31	3.63	94.76
	$\mathcal{L}_2$+$\mathcal{L}_{per}$+$\mathcal{L}_{teg}$	97.77	94.49	99.31	95.70	92.53	2.85	96.02
	$\mathcal{L}_2$+$\mathcal{L}_{per}$+$\mathcal{L}_{teg}$+$\mathcal{L}_{adv}$	96.96	96.68	99.45	96.46	93.85	1.80	96.72

2) Effect of Different Data Type Configurations: This section examines to what extent the introduced polarization characteristics can help LMPC2D3DCNet to detect road regions. Table III summarizes the discriminative performance comparison of these features including S_0, S_0+DoP, S_0+AoP, and S_0+ DoP + AoP. It is observed from Table III that, the performance of S_0+ DoP + AoP consistently transcends other features. The introduction of the DoP and AoP obtains 0.27% 5.44%, 45.37%, and 2.87% average improvements of *OA*, *IoU*, *BER*, and *F1* scores, respectively, compared with S_0on both LDDRS validation and testing datasets. The above comparison results can demonstrate that the impact of feature combination should be positive.

3) Effect of Different CNN Configurations for MPC2D3DCNet: In contrast to previous methods, the underly hypothesis behind our method is that the proposed MPC2D3DCNet can more effectively model the contextual semantic relationship between road and its background, due to that the *MPE* can gradually enrich the dense detailed cues in the *x-y* plane and scene contour cues along the z-axis provided by multiple polarization characteristics. To illustrate this point, we respectively train two modified models: (1) constructing MPE and MPD using pure 2D CNNs, and (2) building MPE and MPD using pure 3D CNNs. Subsequently, the performances of these models are compared to that of the proposed MPC2D3DCNet. Table IV reports the performance

TABLE III

PERFORMANCE COMPARISON OF DIFFERENT DATA TYPE CONFIGURATIONS ON BOTH LDDRS VALIDATION DATASET AND TESTING DATASET, NUMBER HIGHLIGHTED WITH RED TO INDICATE THE BEST RESULTS. ↑&↓ DENOTE LARGER AND SMALLER IS BETTER, RESPECTIVELY

Dataset	Data Type	*PRE*(%) ↑	*REC*(%) ↑	*OA*(%) ↑	*MCC*(%) ↑	*IoU*(%) ↑	*BER*(%) ↓	*F1*(%) ↑
VAL106	S_0	94.10	93.78	99.36	93.53	88.70	3.28	93.77
	S_0+DoP	98.49	91.80	99.23	94.59	90.56	4.17	94.85
	S_0+AoP	96.74	94.98	99.28	95.40	92.02	2.66	95.75
	S_0+DoP+AoP	97.15	95.93	99.41	96.15	93.33	2.17	96.41
TEST317	S_0	97.11	91.21	98.96	93.49	88.82	4.50	93.94
	S_0+DoP	96.94	94.44	99.22	95.19	91.76	2.90	95.55
	S0+AoP	96.79	95.67	99.36	95.82	92.74	2.30	96.10
	S_0+DoP+AoP	97.11	96.38	99.45	96.41	93.84	1.94	96.68

TABLE IV

PERFORMANCE COMPARISON OF DIFFERENT CONFIGURATIONS FOR MPC2D3DCNET ON BOTH LDDRS VALIDATION DATASET AND TESTING DATASET, NUMBER HIGHLIGHTED WITH RED TO INDICATE THE BEST RESULTS. ↑&↓ DENOTE LARGER AND SMALLER IS BETTER, RESPECTIVELY

Dataset	Models	*PRE*(%) ↑	*REC*(%) ↑	*OA*(%) ↑	*MCC*(%) ↑	*IoU*(%) ↑	*BER*(%) ↓	*F1*(%) ↑
VAL106	2D CNN	94.60	92.44	99.32	93.07	87.90	3.92	93.33
	3D CNN	94.10	93.78	99.36	93.53	88.70	3.28	93.77
	Our C2D3DC	96.73	96.18	99.39	96.05	93.17	2.05	96.32
TEST317	2D CNN	98.07	92.17	99.14	94.52	90.52	4.00	94.83
	3D CNN	97.10	93.38	99.10	94.64	90.80	3.42	95.01
	Our C2D3DC	97.49	96.05	99.44	96.42	93.79	2.09	96.69

TABLE V

PERFORMANCE COMPARISON BETWEEN WITH AND WITHOUT C2D3DNLA MODULE ON BOTH LDDRS VALIDATION DATASET AND TESTING DATASET, NUMBER HIGHLIGHTED WITH RED TO INDICATE THE BEST RESULTS. ↑&↓ DENOTE LARGER AND SMALLER IS BETTER, RESPECTIVELY

Dataset	Models	*PRE*(%) ↑	*REC*(%) ↑	*OA*(%) ↑	*MCC*(%) ↑	*IoU*(%) ↑	*BER*(%) ↓	*F1*(%) ↑
VAL106	w/o C2D3DNLA	93.59	94.53	99.33	93.56	88.63	2.90	93.77
	w/ C2D3DNLA	96.76	96.06	99.39	96.00	93.10	2.12	96.27
TEST317	w/o C2D3DNLA	98.24	92.73	99.27	95.00	91.41	3.71	95.30
	w/ C2D3DNLA	97.12	96.53	99.46	96.49	93.97	1.86	96.76

comparison of different models on both LDDRS validation and testing datasets. Specifically, taking the results from the testing datasets as an example, MPC2D3DCNet based on Collaborative 2D/3D CNN achieves significant average *OA*, *IoU*, *BER*, and *F1* boosts of 0.30%, 3.61%, 47.75%, and 1.96% respectively compared to those of 2D CNN, as well as boosts of 0.34%, 3.29%, 38.89%, and 1.77% respectively compared to those of 3D CNN. This indicates that the proposed MPC2D3DCNet method, which considers the relationship among multiple polarization characteristics, sparse inter-channel information, and dense intra-channel information, has broken through the bottleneck of the previous methods, which ignore the unique spatial property of polarization.

4) Effectiveness of C2D3DNLA Module: This purpose of this section is to further experimentally demonstrate the efficacy of our proposed MPC2D3DCNet with and without C2D3DNLA. We report the evaluation results of the proposed MPC2D3DCNet with or without C2D3DNLA on the both LDDRS validation and testing datasets in Tab. V. It can be observed that the performance of MPC2D3DCNet is indeed enhanced after applying C2D3DNLA. Using the results of the testing dataset as an example, the average *OA*, *IoU*, *BER*, and *F1* values of using C2D3DNLA are 99.46%, 93. 97%, 1.86%, and 96.76%, which are 0.19%, 2.80%, 49.87%, and 1.53% higher than those of without using C2D3DNLA.

F. Discussion

1) Analysis of Efficiency: Table VI displays a comparison of *Params* and *FLOPs* of different methods. All state-of-the-art methods are evaluated on 512×640 images with a single Nvidia GeForce RTX 3090 GPU using their publicly available code. As shown in Table VI, we can observe that the CGNet, LRSR-net, and ENet employ fewer *Params*, *i.e.*,

TABLE VI
COMPARISON OF THE COMPUTATIONAL COMPLEXITY OF DIFFERENT METHODS, NUMBER HIGHLIGHTED WITH RED, GREEN, AND BLUE TO INDICATE THE BEST THREE RESULTS

Method	SegNet	LinkNet	UNet	ResUNet	Unet++	UNet3+	CGNet	DABNet
Params (M)	29.44	11.53	31.04	8.22	36.63	26.97	0.49	0.75
FLOPs (G)	200.84	15.15	208.62	228.61	693.32	999.66	4.44	1.32
Method	LRSR-net	ContextNet	ENet	PSPNet	DenseASPP	MSADNet	FastSCNN	ERFNet
Params (M)	0.45	0.87	0.35	65.57	28.6	22.14	1.14	2.06
FLOPs (G)	34.16	1.12	3.6	322.5	154.17	979.23	0.44	18.43
Method	DeepLabv3+(Xception)		DeepLabv3+(ResNet101)		LaneNet	LDNet	DDRNet	LMPC2D3DCNet
Params (M)	54.61		59.33		0.53	5.71	32.32	0.13
FLOPs (G)	103.88		111.19		4.22	81.5	43.92	2.65

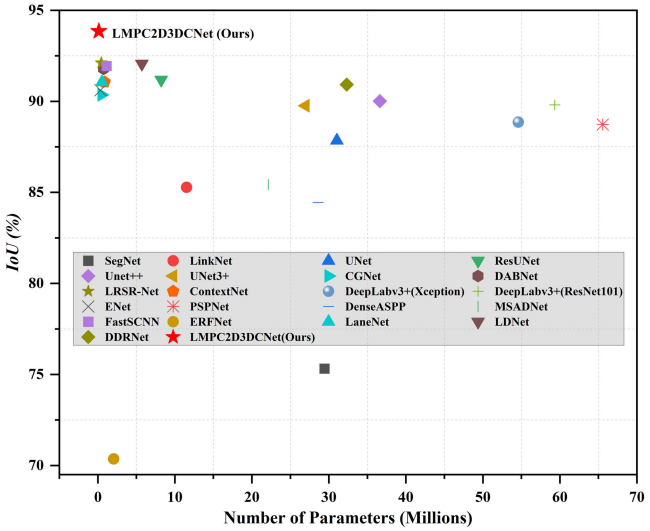

Fig. 13. Intersection over Union (*IoU*) v.s. model parameters. All models are trained on the LWIR DoFP Dataset of Road Scene at resolution 512×640 from scratch.

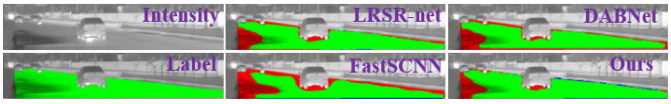

Fig. 14. Failure cases tend to occur on curbsides covered with unexpected objects such as standing water, which could be caused by the limited number of images containing curbside coverings in the training dataset. Green regions, blue regions, and red regions represent true positive, false positive, and false negative, respectively.

0.49M, 0.45M and 0.35M respectively, whereas the *Params* of our LMPC2D3DCNet is only 0.13M (about $\frac{1}{4}$ of CGNet, $\frac{3}{10}$ of LRSR-net, and $\frac{7}{20}$ of ENet). Meanwhile, our method also gives the fourth best performance in terms of *FLOPs*. The underlying factor contributing to this outcome is the excessive computational overhead incurred by the C2D3DNLA. In the subsequent works, we will propose targeted strategies to counter the above issue. Fig. 13 also shows intersection over Union (*IoU*) v.s. model parameters. Our proposed LMPC2D3DCNet strikes a balance between high accuracy and lightweight design, achieving the best of both worlds.

2) Analysis of Limitations: Although LMPC2D3DCNet exhibits great robustness in most cases, when it comes to curbsides covered with unexpected objects (non-object shadows) such as water, LMPC2D3DCNet offers limited superiority to other advanced methods and cannot produce good results, as failure cases shown in Fig. 14. This is expected because there are only few images containing curbside coverings in the training dataset, thus, the *MPE* of LMPC2D3DCNet is unable to learn sufficient knowledge for this case, leading to less effectiveness in road detection.

V. CONCLUSION

In this paper, we propose a novel approach LMPC2D3DCNet designed from the ground up specifically for identifying road regions by combining DL and LWIR-PIs. The LMPC2D3DCNet leverages intensity information, DoP and AoP, along with a novel multi-pathway joint 2D/3D convolutional networks architecture, to establish the correlations among polarization, 2D plane, 3D space, and road region. Extensive experiments on a public infrared polarization dataset of road scenes demonstrate that our proposed LMPC2D3DCNet achieves state-of-the-art performance. However, there are still challenges in effectively utilizing multiple polarization characteristics to enhance full-time road detection, and further advancements are needed in characterizing polarization information and developing light-weight network model. Last but not least, much improvement has been achieved by our proposed LMPC2D3DCNet. We remain committed to making progress in these areas

ACKNOWLEDGMENT

The computation of this research was supported by the HPC Platform of Hefei University of Technology.

REFERENCES

[1] S. B. Powell, R. Garnett, J. Marshall, C. Rizk, and V. Gruev, "Bioinspired polarization vision enables underwater geolocalization," *Sci. Adv.*, vol. 4, no. 4, Apr. 2018, Art. no. eaao6841.

[2] X. Qiao, Y. Zhao, L. Chen, S. G. Kong, and J. Cheung-Wai Chan, "Mosaic gradient histogram for object tracking in DoFP infrared polarization imaging," *ISPRS J. Photogramm. Remote Sens.*, vol. 194, pp. 108–118, Dec. 2022.

[3] N. Li, B. L. Teurnier, M. Boffety, F. Goudail, Y. Zhao, and Q. Pan, "No-reference physics-based quality assessment of polarization images and its application to demosaicking," *IEEE Trans. Image Process.*, vol. 30, pp. 8983–8998, 2021.

[4] Q. Li et al., "LO-Net: Deep real-time LiDAR odometry," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 8465–8474, doi: 10.1109/CVPR.2019.00867.

[5] A. B. Hillel, R. Lerner, D. Levi, and G. Raz, "Recent progress in road and lane detection: A survey," *Mach. Vis. Appl.*, vol. 25, no. 3, pp. 727–745, 2014, doi: 10.1007/s00138-011-0404-2.

[6] A. S. Akopov and L. A. Beklaryan, "Traffic improvement in Manhattan road networks with the use of parallel hybrid biobjective genetic algorithm," *IEEE Access*, vol. 12, pp. 19532–19552, 2024.

[7] H.-F. Wang et al., "Vehicle-road environment perception under low-visibility condition based on polarization features via deep learning," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 10, pp. 17873–17886, Oct. 2022.

[8] S. Aïnouz, J. Zallat, A. de Martino, and C. Collet, "Physical interpretation of polarization-encoded images by color preview," *Opt. Exp.*, vol. 14, no. 13, pp. 5916–5927, 2006.

[9] N. Li, Y. Zhao, Q. Pan, S. G. Kong, and J. C.-W. Chan, "Illumination-invariant road detection and tracking using LWIR polarization characteristics," *ISPRS J. Photogramm. Remote Sens.*, vol. 180, pp. 357–369, Oct. 2021.

[10] K. Li, M. Qi, S. Zhuang, Y. Yang, and J. Gao, "TIPFNet: A transformer-based infrared polarization image fusion network," *Opt. Lett.*, vol. 47, no. 16, pp. 4255–4258, 2022.

[11] N. Li, Y. Zhao, Q. Pan, S. G. Kong, and J. C.-W. Chan, "Full-time monocular road detection using zero-distribution prior of angle of polarization," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Glasgow, U.K. Berlin, Germany: Springer, Aug. 2020. pp. 457–473.

[12] N. Li, Y. Zhao, R. Wu, and Q. Pan, "Polarization-guided road detection network for LWIR division-of-focal-plane camera," *Opt. Lett.*, vol. 46, pp. 5679–5682, Jan. 2021.

[13] A. Paszke, A. Chaurasia, S. Kim, and E. Culurciello, "ENet: A deep neural network architecture for real-time semantic segmentation," 2016, *arXiv:1606.02147*.

[14] M. Yang, K. Yu, C. Zhang, Z. Li, and K. Yang, "DenseASPP for semantic segmentation in street scenes," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3684–3692.

[15] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder–decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.

[16] Y. Qian, J. M. Dolan, and M. Yang, "DLT-Net: Joint detection of drivable areas, lane lines, and traffic objects," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 11, pp. 4670–4679, Dec. 2019.

[17] G. Li, I. Yun, J. Kim, and J. Kim, "DABNet: Depth-wise asymmetric bottleneck for real-time semantic segmentation," 2019, *arXiv:1907.11357*.

[18] X. Lu, "Self-supervised road detection from a single image," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2015, pp. 2989–2993.

[19] M. Shao, C. Xia, Z. Yang, J. Huang, and X. Wang, "Transparent shape from a single view polarization image," 2022, *arXiv:2204.06331*.

[20] A. Chaurasia and E. Culurciello, "LinkNet: Exploiting encoder representations for efficient semantic segmentation," in *Proc. IEEE Vis. Commun. Image Process. (VCIP)*, Dec. 2017, pp. 1–4.

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

[22] Z. Zhang, Q. Liu, and Y. Wang, "Road extraction by deep residual U-Net," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 5, pp. 749–753, May 2018.

[23] 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. (ECCV)*, 2018, pp. 801–818.

[24] D. Liu, J. Zhang, Y. Wu, and Y. Zhang, "A shadow detection algorithm based on multiscale spatial attention mechanism for aerial remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[25] R. P. K. Poudel, U. Bonde, S. Liwicki, and C. Zach, "ContextNet: Exploring context and detail for semantic segmentation in real-time," 2018, *arXiv:1805.04554*.

[26] R. P. K. Poudel, S. Liwicki, and R. Cipolla, "Fast-SCNN: Fast semantic segmentation network," 2019, *arXiv:1902.04502*.

[27] S. Sun, Z. Yang, and T. Ma, "Lightweight remote sensing road detection network," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[28] T. Wu, S. Tang, R. Zhang, J. Cao, and Y. Zhang, "CGNet: A light-weight context guided network for semantic segmentation," *IEEE Trans. Image Process.*, vol. 30, pp. 1169–1179, 2021.

[29] 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.*, 2015, pp. 234–241.

[30] P. Mehrannia, S. S. G. Bagi, B. Moshiri, and O. A. Al-Basir, "Deep representation of imbalanced spatio-temporal traffic flow data for traffic accident detection," *IET Intell. Transp. Syst.*, vol. 17, no. 3, pp. 606–619, Mar. 2023.

[31] P. Wu, T. Chen, Y. Diew Wong, X. Meng, X. Wang, and W. Liu, "Exploring key spatio-temporal features of crash risk hot spots on urban road network: A machine learning approach," *Transp. Res. A, Policy Pract.*, vol. 173, Jul. 2023, Art. no. 103717.

[32] G. Jin et al., "Spatio-temporal graph neural networks for predictive learning in urban computing: A survey," *IEEE Trans. Knowl. Data Eng.*, early access, Nov. 23, 2023, doi: 10.1109/TKDE.2023.3333824.

[33] J. Huang, W. Liu, and X. Sun, "A pavement crack detection method combining 2D with 3D information based on Dempster–Shafer theory," *Comput.-Aided Civil Infrastruct. Eng.*, vol. 29, no. 4, pp. 299–313, Apr. 2014.

[34] Y. Hu and T. Furukawa, "Automatic detection and evaluation of 3D pavement defects using 2D and 3D information at the high speed," *Int. J. Automot. Eng.*, vol. 9, no. 4, pp. 323–331, 2018.

[35] K. Bayoudh, F. Hamdaoui, and A. Mtibaa, "Transfer learning based hybrid 2D-3D CNN for traffic sign recognition and semantic road detection applied in advanced driver assistance systems," *Appl. Intell.*, vol. 51, pp. 124–142, Aug. 2021.

[36] F. Han, T. Mu, H. Li, and A. Tuniyazi, "Deep image prior plus sparsity prior: Toward single-shot full-Stokes spectropolarimetric imaging with a multiple-order retarder," *Adv. Photon. Nexus*, vol. 2, no. 3, May 2023, Art. no. 036009.

[37] B. Lin, X. Fan, D. Li, and Z. Guo, "High-performance polarization imaging reconstruction in scattering system under natural light conditions with an improved U-Net," *Photonics*, vol. 10, no. 2, p. 204, Feb. 2023.

[38] X. Fan, W. Chen, B. Lin, P. Peng, and Z. Guo, "Improved polarization scattering imaging using local–global context polarization feature learning framework," *Opt. Lasers Eng.*, vol. 178, Jul. 2024, Art. no. 108194.

[39] D. Li, B. Lin, X. Wang, and Z. Guo, "High-performance polarization remote sensing with the modified U-Net based deep-learning network," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5621110.

[40] B. Lin, X. Fan, and Z. Guo, "Self-attention module in a multi-scale improved U-Net (SAM-MIU-Net) motivating high-performance polarization scattering imaging," *Opt. Exp.*, vol. 31, no. 2, pp. 3046–3058, 2023.

[41] W. Yu et al., "Polarized computational ghost imaging in scattering system with half-cyclic sinusoidal patterns," *Opt. Laser Technol.*, vol. 169, Feb. 2024, Art. no. 110024.

[42] C. Xu et al., "High-performance deep-learning based polarization computational ghost imaging with random patterns and orthonormalization," *Phys. Scripta*, vol. 98, no. 6, 2023, Art. no. 065011.

[43] X. Fan, B. Lin, K. Guo, B. Liu, and Z. Guo, "TSMPN-PSI: High-performance polarization scattering imaging based on three-stage multi-pipeline networks," *Opt. Exp.*, vol. 31, no. 23, pp. 38097–38113, 2023.

[44] B. Lin, X. Fan, P. Peng, and Z. Guo, "Dynamic polarization fusion network (DPFN) for imaging in different scattering systems," *Opt. Exp.*, vol. 32, no. 1, pp. 511–525, 2024.

[45] H. G. Berry, G. Gabrielse, and A. E. Livingston, "Measurement of the Stokes parameters of light," *Appl. Opt.*, vol. 16, no. 12, pp. 3200–3205, 1977.

[46] Z. Dong et al., "MNet: Rethinking 2D/3D networks for anisotropic medical image segmentation," 2022, *arXiv:2205.04846*.

[47] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7794–7803, doi: 10.1109/CVPR.2018.00813.

[48] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 586–595.

[49] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[50] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1125–1134.

[51] A. Abubakar, X. Zhao, S. Li, M. Takruri, E. Bastaki, and A. Bermak, "A block-matching and 3-D filtering algorithm for Gaussian noise in DoFP polarization images," *IEEE Sensors J.*, vol. 18, no. 18, pp. 7429–7435, Sep. 2018.

[52] N. Li, Y. Zhao, Q. Pan, and S. G. Kong, "Demosaicking DoFP images using Newton's polynomial interpolation and polarization difference model," *Opt. Exp.*, vol. 27, no. 2, pp. 1376–1391, 2019.

[53] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–12.

[54] E. Romera, J. M. Álvarez, L. M. Bergasa, and R. Arroyo, "ERFNet: Efficient residual factorized convnet for real-time semantic segmentation," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 1, pp. 263–272, Jan. 2017.

[55] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: A nested U-Net architecture for medical image segmentation," in *Proc. Int. Workshop Deep Learn. Med. Image Anal. Multimodal Learn. Clin. Decis. Support*, 2018, pp. 3–11.

[56] H. Huang et al., "UNet 3+: A full-scale connected UNet for medical image segmentation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 1055–1059, doi: 10.1109/ICASSP40776.2020.9053405.

[57] D. Neven, B. D. Brabandere, S. Georgoulis, M. Proesmans, and L. V. Gool, "Towards end-to-end lane detection: An instance segmentation approach," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Jun. 2018, pp. 286–291.

[58] F. Munir, S. Azam, M. Jeon, B.-G. Lee, and W. Pedrycz, "LDNet: End-to-end lane marking detection approach using a dynamic vision sensor," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 7, pp. 9318–9334, Jul. 2022.

[59] H. Pan, Y. Hong, W. Sun, and Y. Jia, "Deep dual-resolution networks for real-time and accurate semantic segmentation of traffic scenes," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 3, pp. 3448–3460, Mar. 2023.

[60] J. Cuzick, "A Wilcoxon-type test for trend," *Stat. Med.*, vol. 4, no. 1, pp. 87–90, 1985.

Xueqiang Fan is currently pursuing the Ph.D. degree with the School of Computer Science and Information Engineering, Hefei University of Technology, Hefei, China. His current research interests include pattern recognition and polarization vision.

Bing Lin received the B.E. degree in communication engineering from Hefei University of Technology, Hefei, China, in 2021, where she is currently pursuing the master's degree with the Advanced Electromagnetic Function Laboratory (AEMFLab). Her research interests include polarization imaging and deep learning.

Zhongyi Guo received the bachelor's degree from the Department of Physics, Harbin Institute of Technology, Harbin, China, in 2003, and the master's and Ph.D. degrees from Harbin Institute of Technology in 2005 and 2008, respectively. From 2008 to 2009, he was an Assistant Professor with the Department of Physics, Harbin Institute of Technology. He was a Post-Doctoral Researcher with Hanyang University, Seoul, South Korea, for two years. In 2011, he continued to move to The Hong Kong Polytechnic University, Hong Kong, as a Post-Doctoral Researcher, for six months. In the end of 2011, he joined the School of Computer and Information, Hefei University of Technology, Hefei, China, as a Full Professor. His research interests include polarization information processing, advanced optical communication, OAM antenna, manipulation of optical fields, and nanophotonics.