

Article

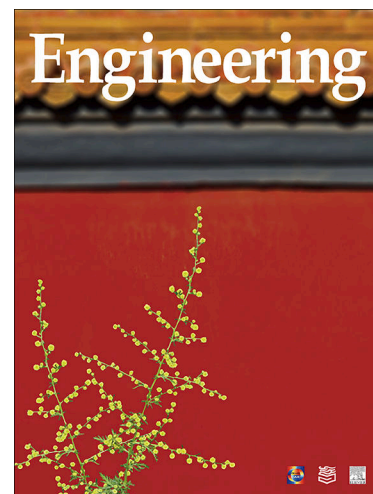
ABCDWaveNet: Advancing Robust Road Ponding Detection in Fog Through Dynamic Frequency-Spatial Synergy

Ronghui Zhang, Dakang Lyu, Tengfei Li, Yunfan Wu, Ujjal Manandhar, Benfei Wang, Junzhou Chen, Bolin Gao, Danwei Wang, Yiqiu Tan

PII: S2095-8099(26)00118-9
DOI: <https://doi.org/10.1016/j.eng.2026.01.026>
Reference: ENG 2273

To appear in: *Engineering*

Received Date: 9 April 2025
Revised Date: 8 November 2025
Accepted Date: 25 January 2026



Please cite this article as: R. Zhang, D. Lyu, T. Li, Y. Wu, U. Manandhar, B. Wang, J. Chen, B. Gao, D. Wang, Y. Tan, ABCDWaveNet: Advancing Robust Road Ponding Detection in Fog Through Dynamic Frequency-Spatial Synergy, *Engineering* (2026), doi: <https://doi.org/10.1016/j.eng.2026.01.026>

This is a PDF of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability. This version will undergo additional copyediting, typesetting and review before it is published in its final form. As such, this version is no longer the Accepted Manuscript, but it is not yet the definitive Version of Record; we are providing this early version to give early visibility of the article. Please note that Elsevier's sharing policy for the Published Journal Article applies to this version, see: <https://www.elsevier.com/about/policies-and-standards/sharing#4-published-journal-article>. Please also note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

—Article

ABCDWaveNet: Advancing Robust Road Ponding Detection in Fog Through Dynamic Frequency-Spatial Synergy

Ronghui Zhang ^a, Dakang Lyu ^a, Tengfei Li ^a, Yunfan Wu ^a, Ujjal Manandhar^{b,c}, Benfei Wang ^a, Junzhou Chen ^{a,*}, Bolin Gao ^d, Danwei Wang ^b, Yiqiu Tan ^e

^a Guangdong Provincial Key Laboratory of Intelligent Transportation System, School of Intelligent Systems Engineering, Sun Yat-sen University, Guangzhou 510275, China

^b School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 639798, Singapore

^c Western Power, Perth WA 6000, Australia

^d School of Vehicle and Mobility, Tsinghua University, Beijing 150090, China

^e School of Transportation Science and Engineering, Harbin Institute of Technology, Harbin 10084, China

*Corresponding author.

E-mail address: chenjunzhou@mail.sysu.edu.cn (J. Chen).

Abstract: Road ponding presents a substantial threat to vehicular safety, particularly in foggy conditions where reliable detection continues to be a major challenge for advanced driver assistance systems (ADASs). To address this issue, we propose an aggregation-broadcast-coupling dynamic wavelet network (ABCDWaveNet), a novel deep learning framework specifically designed to achieve robust ponding detection in fog-affected environments. The central architecture of ABCDWaveNet improves detection performance by utilizing dynamic convolution for adaptive feature extraction under reduced visibility, together with a wavelet-based module that improves feature representation across both spatial and frequency domains, thereby effectively alleviating fog-related interference. In addition, ABCDWaveNet incorporates multi-scale structural and contextual information and employs an adaptive attention coupling gate to dynamically integrate global and local features, leading to improved detection accuracy. For realistic evaluations under compounded adverse weather conditions, we introduce the Foggy Low-Light Puddle dataset. Comprehensive experiments confirmed that ABCDWaveNet attained state-of-the-art results, with notable intersection over union gains of 3.51%, 1.75%, and 1.03% on the Foggy-Puddle, Puddle-1000, and Foggy Low-Light Puddle datasets, respectively. Furthermore, with an inference speed (FPS) of 25.48 on the NVIDIA Jetson AGX Orin, the proposed framework demonstrates strong suitability for development in ADAS applications. These results highlight the effectiveness of ABCDWaveNet, presenting valuable advancements for proactive road safety under challenging weather conditions.

Keywords: Road ponding detection; Advanced driver assistance systems; Deep learning; Dynamic convolution; Wavelet transform; Foggy conditions

1. Introduction

In modern transportation systems, the rapid growth in traffic volume driven by ongoing urbanization has increased the need to ensure road safety in all-weather environments. Although substantial progress has been made in vehicle safety technologies and traffic infrastructure, ensuring safe driving under adverse weather conditions, particularly fog, continues to be a major concern. Fog is among the most frequently occurring weather events and can significantly degrade visibility, which is essential for safe vehicle operation [1–**Error! Reference source not found.**].

Under foggy conditions, drivers face difficulty accurately perceiving their surroundings, including nearby vehicles and pedestrians, as well as potential roadway hazards, such as ponding and potholes.

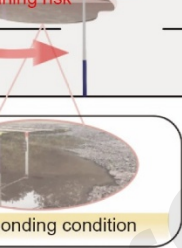
This reduction in visibility substantially amplifies the risks posed by such hazards. For example, under normal driving conditions, standing water on the roadway can cause vehicles to skid and increase braking distances. In foggy environments, these risks are further exacerbated, as drivers increasingly struggle to detect road ponding in time, which can lead to serious consequences for life and property [4,5]. According to the Federal Highway Administration, more than 38 700 vehicle accidents occur annually under foggy conditions in the United States, resulting in over 600 fatalities and in excess of 16 300

injuries **Error! Reference source not found.** Research suggests that the likelihood of accidents in fog is up to 35-fold greater. Furthermore, these conditions contribute to about 38.3% of traffic congestion and 23% of non-recurring delays, resulting in billions of dollars in economic losses [7,8].

To mitigate these challenges, advanced driver assistance systems (ADASs) have been developed to integrate diverse traffic-related information and support safer, more efficient, and environmentally sustainable mobility. Fig. 1 presents an example of a road ponding detection process within an ADAS framework under foggy conditions **Error! Reference source not found.** Although traditional sensor-based methods for detecting road ponding, such as light detection and ranging (LiDAR) **Error! Reference source not found.**, hydrogel sensors **Error! Reference source not found.**, and infrared devices **Error! Reference source not found.**, perform effectively in clear environments, they exhibit substantial limitations in fog. LiDAR-based approaches, including those applied by Zhao et al. **Error! Reference source not found.**, utilize histogram analysis of point cloud data to identify water-related hazards. However, dense fog degrades LiDAR accuracy by reducing point cloud density and introducing substantial noise, thereby impairing reliable detection. Similarly, hydrogel **Error! Reference source not found.** and infrared sensors **Error! Reference source not found.** are highly vulnerable to signal distortion caused by fog, further restricting their effectiveness under low-visibility conditions. Beyond these performance issues, sensor-based solutions often suffer from limited spatial coverage, high deployment cost, and operational complexity, which challenge their practicality for large-scale implementation.

Machine learning-based methods have also been employed for road ponding detection, utilizing image data to identify hazardous water conditions. Classical approaches, such as support vector machines (SVMs) **Error! Reference source not found.** and random forests **Error! Reference source not found.**, detect wet road surfaces by extracting hand-crafted features including color and texture. For example, the integration of optical flow with SVMs has been applied to identify wet roads **Error! Reference source not found.** However, these methods often struggle to generalize in dynamically changing environments, particularly in fog, where reduced visibility and degraded image quality hinder the extraction of fine-grained discriminatory features **Error! Reference source not found.** Additionally, Adugna et al. [26] demonstrated that although RF achieved better performance than SVMs in large-scale land cover classification with high-dimensional data, it remained insufficient when addressing more complex scenarios.

In contrast, deep learning has emerged as an effective solution across multiple domains, especially for visual analysis and robust feature representation **Error! Reference source not found.** The capacity of deep learning models, such as convolutional neural networks (CNNs) and transformers **Error! Reference source not found.**, which learn hierarchical features, are particularly effective for complex visual tasks. Models such as U-Net **Error! Reference source not found.** and AGSENet **Error! Reference source not found.**, which rely on fixed kernels, have performed well in controlled conditions but have difficulty adapting to the varying visual distortions introduced by fog. In addition, models such as SWNet **Error! Reference source not found.**, which utilize multi-scale fusion and attention mechanisms, still experience performance degradation in foggy conditions due to reduced contrast and the loss of fine details. Moreover, transformer-based models, such as ViT **Error! Reference source not found.** and SeaFormer **Error! Reference source not found.**, show potential in capturing global context but are often computationally inefficient, which limits their practical deployment in real-time systems, such as ADAS **Error! Reference source not found.**



onding condition

(1) **Novel framework design:** We present ABCDWaveNet, a pioneering deep learning architecture that merges dynamic convolution with advanced feature fusion to boost road ponding detection under foggy conditions. In contrast to traditional convolutional architectures that rely on fixed kernels, ABCDWaveNet leverages dynamic convolution to improve adaptability to non-uniform visual patterns arising from varying environmental conditions. By simultaneously capturing spatial and frequency domain features, the framework substantially improves robustness and accuracy in challenging foggy environments.

(3) Multi-scale feature integration: The proposed ABC mechanism integrates multi-scale feature aggregation with an adaptive attention coupling gate (AACG), enabling effective fusion of global and local representations. By dynamically modulating the information exchange between scales, this mechanism supports adaptability to varying ponding extents while maintaining high segmentation accuracy in the decoding process.

2. Related work

2.1. Sensor-based methods

Error! Reference source not found.,Error! Reference source not found., are commonly applied for detecting water on road surfaces, providing reliable performance across different environmental settings. LiDAR-based techniques typically perform real-time segmentation and hazard identification by analyzing the point cloud histogram **Error! Reference source not found.** Similarly, near-infrared optoelectronic techniques utilize reflection, scattering, and polarization characteristics to differentiate road surface conditions **Error! Reference source not found..**

Despite these strengths, significant drawbacks arise under foggy-affected scenarios. Fog-related signal attenuation markedly degrades the performance of hydrogel sensors **Error! Reference source not found.** and infrared devices **Error! Reference source not found.**, and reduced visibility compromises the accuracy of LiDAR systems **Error! Reference source not found.** Moreover, the relatively narrow detection range of these sensors, together with the substantial costs and intricate hardware requirements, limits their feasibility for widespread deployment in challenging weather environments.

2.2. Machine learning-based methods

2.2.1. Classical machine learning methods

Classical machine learning techniques, such as SVMs and random forests, have been employed for water detection using hand-crafted descriptors, including color and texture cues. Approaches integrating optical flow with SVMs **Error! Reference source not found.** and SVM-based hypothesis generation relying on color characteristics **Error! Reference source not found.** have been developed to identify wet road surfaces. In addition, optimized SVMs combined with image segmentation **Error! Reference source not found.**, **Error! Reference source not found.** and random forest classifiers utilizing Hue- Saturation-Value (HSV) color features **Error! Reference source not found.** have demonstrated satisfactory performance in controlled environments. Wavelet packet transform-enhanced SVM models have also been introduced to improve wet surface recognition **Error! Reference source not found.**. Nevertheless, although these techniques perform well in simplified scenarios, they exhibit high sensitivity to noise and reduced adaptability in complex, dynamically changing environments, particularly under fog, where visual degradation undermines the reliability of hand-crafted features **Error! Reference source not found.**.

2.2.2. Deep learning-based methods

Deep learning techniques, particularly CNNs, have significantly advanced road water detection and now achieve SOTA performance **Error! Reference source not found.** U-Net **Error! Reference source not found.**, characterized by its encoder-decoder architecture, has become a core framework for segmentation and consistently surpasses classical HSV-based approaches **Error! Reference source not found.** Enhanced variants, such as U-Net-RAU **Error! Reference source not found.**, exploit water reflection characteristics to improve pixel-level predictions, while SWNet **Error! Reference source not found.** integrates multiscale feature fusion with splash attention modules for more accurate water splash recognition. AGSENet **Error! Reference source not found.** incorporates multi-scale attention to strengthen feature discrimination, enabling more robust detection in complex environmental conditions.

Transformer-based approaches, including vision Transformers **Error! Reference source not found.**, have further broadened capabilities in semantic segmentation. SeaFormer **Error! Reference source not found.**, a lightweight Transformer model, effectively captures both global contextual information and local details with efficiency. However, these models still encounter constraints in computational cost and boundary precision, both of which are crucial for fine-grained water segmentation.

2.3. Foggy condition detection

Foggy environments intensify the challenges of ponding detection due to reduced contrast and visibility **Error! Reference source not found.** Classical dehazing techniques, such as the dark channel prior **Error! Reference source not found.**, enhance sense clarity but often degrade fine details, which limits their suitability for complex segmentation tasks. Deep models, such as AOD-Net **Error! Reference source not found.**, provide real-time dehazing but are not tailored for water segmentation. More recent strategies focus on domain adaptation to mitigate fog-induced feature corruption. Fog-invariant features foggy scene segmentation method (FIFO) **Error! Reference source not found.** enhances consistency between clear and foggy domains using fog-pass filtering, and CuDA-Net **Error! Reference source not found.** improves generalization by disentangling fog-related and style-related components. Bi-directional wavelet guidance **Error! Reference source not found.** employs wavelet-based feature decoupling to preserve important details and improve robustness, while its bi-directional design focuses on both global and local context, achieving competitive results in fog-heavy scenarios.

Although these approaches provide progress toward managing fog-related degradation, additional research is required to more tightly couple dehazing mechanisms with task-specific segmentation networks, enabling reliable road ponding detection in adverse weather conditions.

In this section, we begin with an overall description of the network architecture, followed by a detailed explanation of each constituent module.

3.1. Overall architecture

Fig. 2 presents the overall architecture of the proposed ABCDWaveNet. The network adopts a hierarchical U-Net architecture with skip connections to support effective information flow.

Given an red-green-blue (RGB) input image $I \in \mathbb{R}^{H \times W \times C}$ (where $H \times W$ represents spatial resolution, and C indicates the number of channels), ABCDWaveNet begins feature extraction through multiple encoding stages incorporating a dual dynamic convolutional (DDC) layer and a bidomain information synergy (BIS) module. These modules collaboratively generate the encoded feature representations. The features extracted at different encoding levels are subsequently integrated through the multi-scale information aggregation (MIA) module, enabling a thorough multi-scale representation. Operating in a coarse-to-fine manner, the MIA module employs an adaptive scale selection (ASS) mechanism to effectively capture critical structural and contextual cues for ponding detection in foggy conditions. This includes identifying ponding boundaries and distinguishing them from surrounding dry road surfaces. A progressive refinement module is then applied to highlight the most discriminative features, which are essential for separating ponding from non-ponding areas based on variations in color and texture. The globally aggregated feature from the MIA module is propagated to each decoder stage through the AACG embedded within the skip connections. The AACG selectively fuses global and local information, allowing the model to accentuate structure relevance and context-dependent details. In the following subsections, we delve into the details of the newly introduced components.

3.2. Dynamic feature extraction

3.2.1. Dynamic convolution

Extracting reliable features for accurate ponding detection is highly challenging under adverse environments, such as fog. Fog substantially reduces image contrast and blurs edge structures, leading to a noticeable decline in the effectiveness of conventional CNNs **Error! Reference source not found.,Error! Reference source not found.,Error! Reference source not found..** These models often fail to capture essential visual cues under such degraded conditions. To overcome these drawbacks, we propose the incorporation of a double dynamic convolution layer within the encoder, aimed at improving the network's adaptability and robustness through dynamic feature extraction.

The double dynamic convolution layer relies on dynamic convolution, which increases the expressive capability of the network by adaptively combining multiple convolution kernels according to the input content **Error! Reference source not found..** As illustrated in Fig. 2(c), dynamic convolution fuses K parallel convolutional kernels $\{W(k)\}_{k=1}^K$ with input-dependent attention weights (k). For an input feature map $X \in \mathbb{R}^{H \times W \times C}$, the output is computed as

$$Y = \sum_{k=1}^K \alpha^{(k)} \cdot (X * W^{(k)}) \quad (1)$$

where $*$ represents the convolution operation. The coefficient $\alpha^{(k)}$ is adaptively computed from the input via an MLP module. For a given input X , global average pooling is applied to condense the spatial information into a vector. This vector is subsequently passed through the MLP module, followed by a Sigmoid activation function to generate the input-dependent weighting coefficients:

$$\alpha = \text{Sigmoid}(\text{MLP}(\text{Pool}(X))) \quad (2)$$

where $\text{Pool}(X)$ is the result of applying global average pooling to the input feature map X .

Complexity analysis: The parameter count and floating-point operations (FLOPs) of dynamic convolution are evaluated in comparison with those of standard convolution **Error! Reference source not found..** In a standard convolution layer, a single kernel is used, leading to a total parameter count of $C_{\text{out}} \cdot C_{\text{in}} \cdot K \cdot K$, and the corresponding FLOPs can be expressed as $\text{FLOPs}_{\text{conv}} = H' \cdot W' \cdot C_{\text{out}} \cdot C_{\text{in}} \cdot K \cdot K$ (3)

where H' and W' denote the height and width of the output feature map, respectively, C_{out} is the number of output channels, C_{in} is the number of input channels.

In contrast, dynamic convolution incorporates multiple kernels along with an attention mechanism. The total parameter count is determined as $C_{\text{in}}^2 + C_{\text{in}} \cdot M + M \cdot C_{\text{out}} \cdot C_{\text{in}} \cdot K^2$, and the FLOPs for dynamic convolution are expressed as

$$\text{FLOPs} = C_{in}^2 + C_{in}M + MC_{out}C_{in}K^2 + H'W'C_{out}C_{in}K^2(4)$$

Journal Pre-proofs

where M denotes the number of parallel convolutional kernels in the dynamic convolution module.

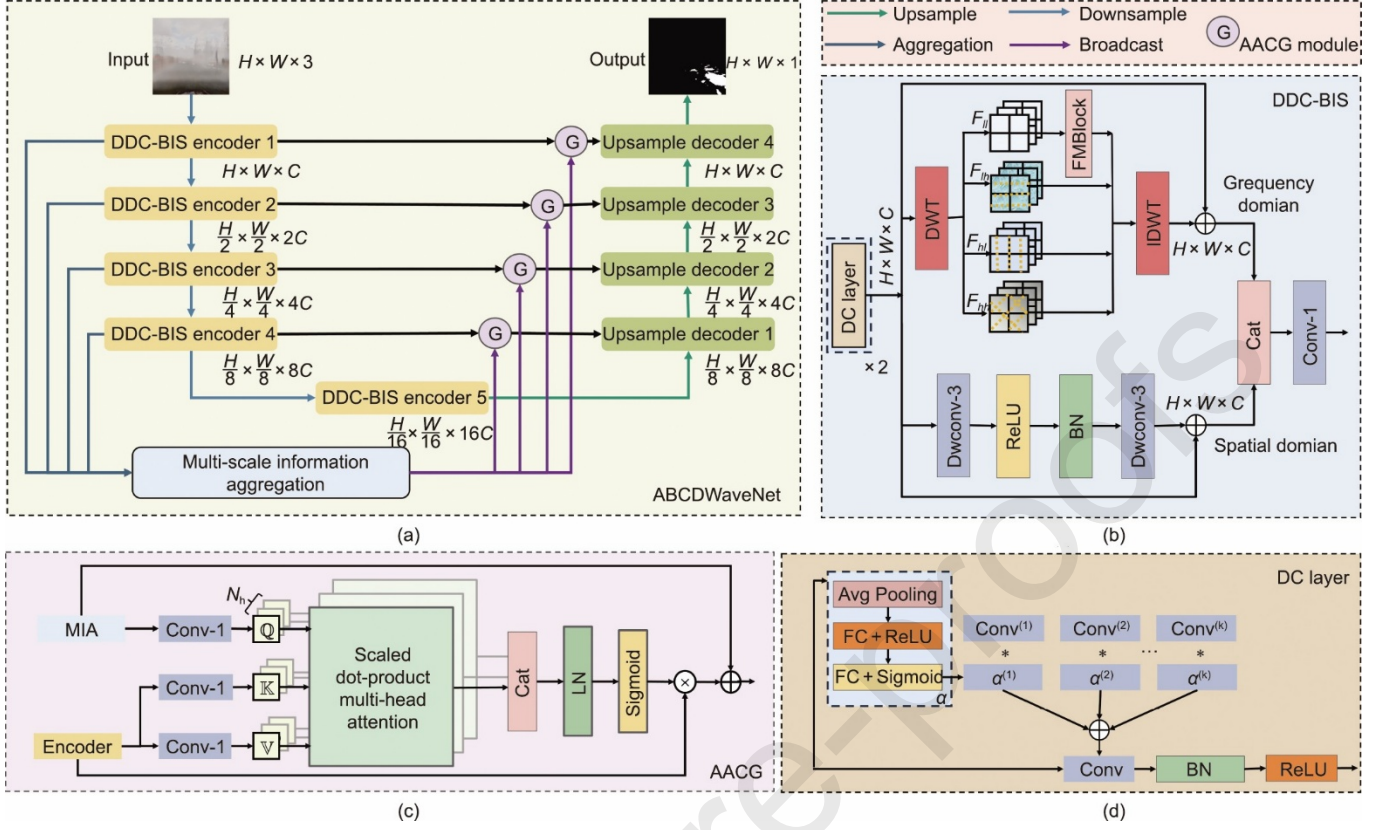


Fig. 2. Architecture of the proposed ABCDWaveNet. (a) A U-shaped network consisting of five encoder stages and four decoder stages, integrated with the MIA and AACG modules. (b) The encoder-decoder framework is constructed by stacking dual dynamic convolution-BIS (DDC-BIS) modules, where (c) illustrates the placement of the MIA and AACG components. (d) Each DDC block contains two sequential dynamic convolution layers. $F_{ll}, F_{lh}, F_{hl}, F_{hh}$: four sub-bands obtained by DWT decomposition of the feature map, where F_{ll} denotes the low-frequency approximation sub-band that preserves the global structural context, and F_{lh}, F_{hl} , and F_{hh} represent the high-frequency detail sub-bands corresponding to the horizontal, vertical, and diagonal directions of the feature map, respectively; $\alpha, \alpha^{(1)}, \alpha^{(2)}, \dots, \alpha^{(k)}$: input-dependent adaptive attention weighting coefficients for the k parallel convolutional kernels in the dynamic convolution module; N_h : number of attention heads in the multi-head cross-attention mechanism; $\mathbf{Q}, \mathbf{K}$, and $\mathbf{V}$: query, key, and value matrices in the multi-head cross-attention mechanism, respectively; BN: batch normalization; LN: layer normalization; FC: fully connected layer; Conv: convolutional layer; ReLU: rectified linear unit.

The parameter ratio of dynamic convolution over standard is

$$\begin{aligned} R_{\text{param}} &= \frac{C_{in}^2 + C_{in}M + MC_{out}C_{in}K^2}{C_{out}C_{in}KK} \\ &= \frac{C_{in}}{C_{out}K^2} + \frac{M}{C_{out}K^2} + (5) \\ &\approx \frac{1}{K^2} + M, \quad (M \ll C_{out}K^2, C_{in} \approx C_{out}) \end{aligned}$$

The FLOP ratio is

$$\begin{aligned} R_{\text{flop}} &= \frac{C_{in}^2 + C_{in}M + MC_{out}C_{in}K^2 + H'W'C_{out}C_{in}K^2}{H'W'C_{out}C_{in}K^2} \\ &= \frac{C_{in}}{H'W'C_{out}K^2} + \frac{M}{H'W'C_{out}K^2} + \frac{M}{H'W'} + (6) \\ &\approx 1, \quad (1 < M \ll H'W', C_{in} \approx C_{out}) \end{aligned}$$

3.2.2. Dual dynamic convolution layer

To enhance the model's adaptability to non-uniform visual patterns arising from diverse environmental conditions, particularly under fog, we propose the DDC layer. This layer enhances the feature extraction process in complex and variable scenes by substituting the conventional pair of fixed convolutional layers, commonly used in architectures such as the U-Net, with dynamic convolutions.

In standard U-Net encoders, feature extraction is performed through two successive static convolutional layers. Although these layers progressively refine feature representations, their fixed nature restricts adaptability to changing input characteristics. By integrating dynamic convolutions, the DDC layer allows the network to modulate its convolution kernels in response to input variations. This adaptive kernel adjustment provides a notable advantage in adverse conditions, such as fog, where reduced visibility and obscured fine structures hinder reliable feature capture.

Formally, the output Y_1 from the first dynamic convolution is computed as

$$Y_1 = \sum_{k=1}^K \alpha_1^{(k)} \cdot (X * W_1^{(k)}) \quad (7)$$

$$\alpha_1^{(k)} = \text{Sigmoid}(\text{MLP}(\text{Pool}(X))) \quad (8)$$

where X denotes the input feature map, and $\alpha_1^{(k)}$ represents the input-dependent attention coefficients as defined in Eq. (8). $W_1^{(k)}$ represent the convolutional kernels of the first layer.

The output Y_2 from the second dynamic convolution is expressed as

$$Y_2 = \sum_{k=1}^K \alpha_2^{(k)} \cdot (Y_1 * W_2^{(k)})$$

$$\alpha_2^{(k)} = \text{Sigmoid}(\text{MLP}(\text{Pool}(Y_1)))$$

where $W_2^{(k)}$ denotes the convolutional kernels of the second layer, and $\alpha_2^{(k)}$ represents the attention coefficients computed from Y_1 as described in Eq. (10).

This dynamic convolution process enables iterative refinement of feature extraction, allowing the model to adaptively emphasize salient features under diverse conditions. The DDC layer enhances robustness in visually degraded environments, such as fog, where image contrast and clarity are reduced. By dynamically adjusting the convolution weights, the model retains essential information, enabling accurate detection of subtle features, such as road and ponding boundaries, even under obscured visibility.

Unlike conventional kernel-fixed convolutions that employ two sequential static layers, dynamic convolutions provide improved flexibility by modulating kernels according to input variation, thereby enhancing feature extraction. Although the parameter count increases slightly, the model sustains computational efficiency with only a marginal rise in FLOPs, as confirmed by the complexity analysis. This makes it particularly suitable for applications such as road safety early-warning systems, where both accuracy and processing speed are crucial.

3.3. Wavelet-based synergy of frequency and spatial information

Detecting ponding under foggy conditions is particularly challenging due to reduced visibility, light scattering, and the reflective nature of water surfaces. These factors significantly diminish image clarity and detail, making it difficult for conventional image processing techniques to accurately identify waterlogged regions. To overcome these limitations, we propose the BIS module, a novel framework designed to enhance detection performance in fog-affected environments.

The BIS module consists of two complementary pathways: the frequency-aware information processing pathway and the spatial-enhanced feature learning pathway. The frequency pathway operates on low-frequency components to preserve global structural consistency while simultaneously exploiting high-frequency components to sharpen edges and recover fine details. This dual representation enables accurate modeling of both the smooth and detailed characteristics of water surfaces partially obscured by fog. In parallel, the spatial pathway focuses on object localization and contextual relationships, facilitating precise identification of ponding regions and their interactions with surrounding features. By integrating both

3.3.1. Frequency-aware information processing pathway

Foggy conditions introduce considerable difficulties in ponding detection by reducing visibility and degrading edge information that is essential for differentiating waterlogged regions. These effects lead to a notable loss of fine structural detail, particularly in the high-frequency components of the image. To investigate the influence of fog on both the frequency and spatial aspects of ponding detection, we employ the DWT using the Haar wavelet basis, selected for its computational simplicity. This decomposition separates the image into four sub-bands (see Fig. 3): F_{LL} , F_{LH} , F_{HL} , and F_{HH} , which represent the low- and high-frequency components. The LL sub-band mainly encodes the global structural context and remains relatively stable under foggy conditions, showing minimal variation between clear and fog-affected scenes. In contrast, the high-frequency sub-bands (LH, HL, and HH) capture critical edge and texture information but experience substantial attenuation as fog density increases, leading to a pronounced loss of detail, especially in regions containing ponding. The results of this analysis are presented in Fig. 4, where the x -axis corresponds to different fog intensities, from clear to slight fog, medium fog, and heavy fog **Error! Reference source not found..** The y -axis denotes the normalized energy ratio for each frequency sub-band (LL, LH, HL, and HH). The results indicated a clear trend: under clear conditions, high-frequency sub-bands (LH, HL, and HH) maintained a greater proportion of the total energy. As fog concentration increased, the energy distribution progressively shifted, with the low-frequency energy (LL) increasing while the high-frequency energy (LH, HL, HH) consistently decreasing.

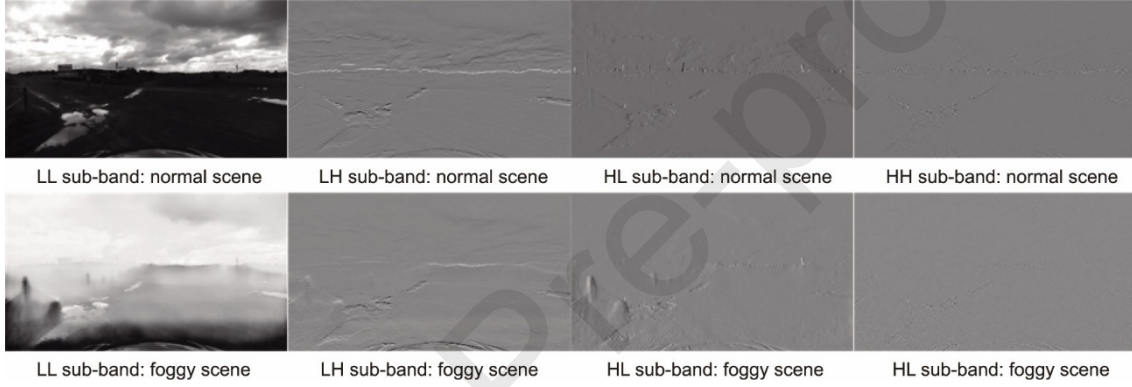


Fig. 3. Wavelet sub-band analysis of Puddle-1000 images and synthesized Foggy-Puddle images.

The frequency-aware information processing pathway aims to reduce the adverse effects of fog while maintaining essential scene details, thereby improving the accuracy of ponding area detection. As shown in Fig. 6, Fig. 6(b), this pathway is composed of three primary components: the feature mixing block (FM-Block) **Error! Reference source not found.**, the DWT, and the inverse DWT (IDWT) **Error! Reference source not found..** For a given input feature map $F \in \mathbb{R}^{H \times W \times D}$ (where D is denotes the number of channels), the DWT decomposes it into four sub-bands: F_{LL} , F_{LH} , F_{HL} , and F_{HH} , each of size $\mathbb{R}^{\frac{H}{2} \times \frac{W}{2} \times D}$; the sub-band F_{LL} corresponds to the low-frequency (approximation) component that preserves the global structural context, whereas F_{LH} , F_{HL} , and F_{HH} capture the high-frequency components containing edge and texture information crucial for detailed scene representation **Error! Reference source not found..**

The DWT utilizes low-pass (l) and high-pass (h) filters defined as

$$l = \frac{1}{\sqrt{2}}[1,1], \quad h = \frac{1}{\sqrt{2}}[1,1](11)$$

These filters extract both smooth (low-frequency) and detailed (high-frequency) components, ensuring that the essential scene information is preserved. The corresponding sub-bands are computed as

$$F_{LL}(u,v) = \sum_{m=0}^{H-1} \sum_{n=0}^{W-1} F(m,n) \cdot l(m-2u) \cdot l(n-2v)(12)$$

$$F_{LH}(u,v) = \sum_{m=0}^{H-1} \sum_{n=0}^{W-1} F(m,n) \cdot l(m-2u) \cdot h(n-2v)(13)$$

$$F_{HH}(u,v) = \sum_{m=0}^{H-1} \sum_{n=0}^{W-1} F(m,n) \cdot h(m-2u) \cdot h(n-2v) \quad (15)$$

where m and n denote the row and column indices of F , respectively, and u and v represent the row and column indices of the sub-bands, respectively. The downsampling by a factor of 2 reduces the spatial resolution of the sub-bands while preserving the essential information.

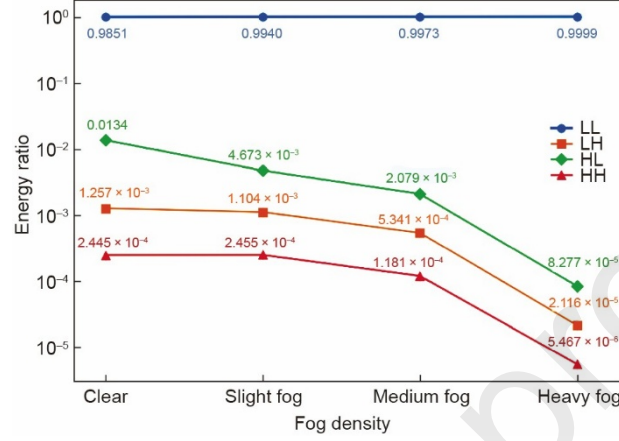


Fig. 4. Energy distribution across frequency bands for different fog density levels.

Unlike conventional wavelet-based approaches **Error! Reference source not found.**, which treat low- and high-frequency components uniformly, our method distinguishes their unique characteristics and processes them separately. The low-frequency sub-band F_{LL} represents the global structural context of the image, which remains relatively stable even under foggy conditions. To enhance its representational capacity while maintaining global coherence, F_{LL} is processed using FM-Block. The FM-Block architecture integrates channel splitting and shuffling (as illustrated in Fig. 5(d)) with FMBCConv (Fig. 5(b)) to facilitate effective feature interaction while preserving both local and global information. This channel splitting and shuffling strategy reduces model complexity by handling smaller subsets of features before recombination. Moreover, it strengthens the network's representational ability by promoting richer inter-channel interactions. The final output after processing F_{LL} through the FM-Block is expressed as

$$\tilde{F}_{LL}^1 = \text{CP}(\text{Conv}(\text{CP}(F_{LL}))) \quad (16)$$

$$\tilde{F}_{LL}^2 = \text{CP}(\text{Conv}(\text{CP}(\tilde{F}_{LL}^1))) \quad (17)$$

$$\hat{F}_{LL} = \text{Conv}(\text{SiLU}(\text{GConv}(\tilde{F}_{LL}^2 + F_{LL}))) \quad (18)$$

where $\text{CP}(\cdot)$ denotes the channel projection (see Fig. 5(d)), $\text{GConv}(\cdot)$ represents the group convolution, and CP corresponds to the channel projection mechanism. SiLU is the Sigmoid linear unit; F_{LL}^1 : first-stage intermediate feature map of the original low-frequency sub-band F_{LL} ; F_{LL}^2 is the second-stage intermediate feature map; $\hat{F}_{LL}$ is the final enhanced low-frequency feature map output; and $\text{Conv}(\cdot)$ represents the convolution.

The enhanced low-frequency sub-band $\hat{F}_{LL}$ and the high-frequency sub-bands are then integrated using IDWT to reconstruct the refined feature map:

$$\hat{F} = \text{IDWT}(\hat{F}_{LL}, F_{LH}, F_{HL}, F_{HH}) \quad (19)$$

where $\hat{F} \in \mathbb{R}^{H \times W \times D}$ combines both global and local information. Finally, a residual connection fuses the original input F with $\hat{F}$, preserving the original structural information while incorporating enriched features:

$$F_{\text{out, frequency}} = F + \hat{F} \quad (20)$$

This design ensures robust and adaptive detection in challenging conditions such as foggy environments.

The spatial-enhanced feature learning pathway enhances spatial feature representation through a series of specialized operations, including depth-wise convolutions, activation functions, and normalization layers. This structure enables the extraction and refinement of local spatial features, providing essential details that complement the global structural information obtained from the frequency pathway.

As shown in Fig. 2(b), the process begins with a depth-wise convolution designed to efficiently capture localized spatial patterns. A ReLU activation function follows, introducing non-linearity and allowing the network to learn complex, fine-grained representations. Batch normalization (BatchNorm) is subsequently applied to stabilize feature distributions, promote training convergence, and enhance overall robustness. A second depth-wise convolutional layer is then employed to further enhance the extraction of spatial features (F'), strengthening the pathway's capability to delineate object boundaries and capture intricate spatial relationships.

$$F' = \mathcal{D}_2(\mathcal{B}(\phi(\mathcal{D}_1(F))))(21)$$

where $\mathcal{D}_1$ and $\mathcal{D}_2$ represent the depth-wise convolution operations, $\phi(\cdot)$ denotes the ReLU activation function, and $\mathcal{B}(\cdot)$ refers to batch normalization.

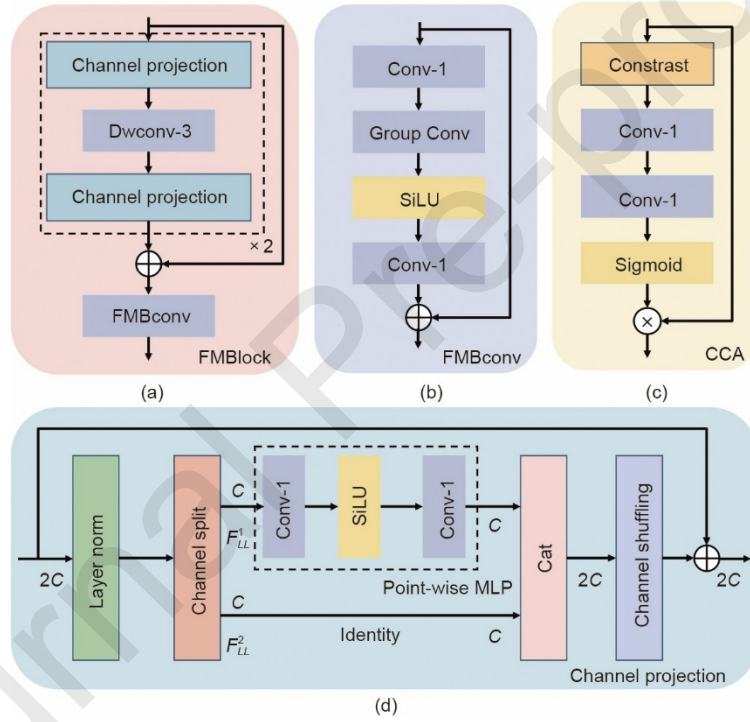


Fig. 5. (a) Architecture of the FM-Block. (b) Architecture of the FM-Block conv (FMBconv). (c) Architecture of contrast-aware channel attention (CCA). (d) Architecture of channel projection. C represents the number of channels. DWconv: Depth-wise Convolution.

To integrate the enriched spatial features with the original input seamlessly, a residual connection is incorporated as follows:

$$F_{\text{out, spatial}} = F' + (22)$$

This integration ensures that the model retains the original information while embedding the enhanced spatial details. This balanced approach strengthens the overall feature representation, enabling precise ponding detection under challenging conditions, such as foggy environments, where maintaining both local and global information is critical.

After processing through both the frequency-aware information processing pathway and the spatial-enhanced feature learning pathway, the outputs are concatenated and refined using a depth-wise separable convolution to achieve cohesive feature integration:

$$F_{\text{final}} = \mathcal{D}_3(\text{Concat}(F_{\text{out, frequency}}, F_{\text{out, spatial}}))(23)$$

where $\text{Concat}(\cdot)$ denotes the concatenation operation, D_n represents the depth-wise separable convolution, and $F_{\text{out, frequency}}$ and $F_{\text{out, spatial}}$ represent the frequency and spatial domain feature maps, respectively. This synergistic process fuses information from both the frequency and spatial domains, enhancing the model's ability to capture comprehensive feature representations. The resulting feature map F_{final} effectively balances global structural context with local fine details, thereby ensuring robust performance in complex and challenging foggy environments.

3.4. ABC mechanism

To overcome the challenge of detecting ponding at multiple spatial scales, we introduce the ABC mechanism. This mechanism comprises two primary components: the MIA and AACG modules. The ABC mechanism aggregates features from different levels and broadcasts global contextual information to each level, enabling effective integration of features across multiple ponding scales while selectively emphasizing crucial detection cues. The subsequent sections provide a detailed description of these two modules.

3.4.1. MIA module

The MIA module, illustrated in Fig. 6(a), serves as a fundamental component of the ABC mechanism and is specifically designed to address the challenges of detecting ponding on fog-affected roads. In such environments, ponding regions are influenced by reduced visibility, complex road structures, and irregular water distribution. To accurately detect ponding under these conditions, it is essential to extract features at multiple spatial scales, enabling the differentiation between both broad contextual patterns and fine-grained details.

The MIA module integrates two key processes: ASS, as shown in Fig. 6(b), and progressive separable refinement (PSR), as shown in Fig. 6(c). These components work in conjunction to capture, refine, and enhance multi-scale features, producing contextually aware representations that significantly improve detection accuracy.

Initially, feature maps from all hierarchical levels, denoted as $\{F_1, F_2, F_3, F_4, F_5\}$, are concatenated. Adaptive average pooling or bi-linear interpolation is applied to align the number of channels across these feature maps with that of F_4 . A subsequent 1×1 convolution is then performed to reduce the dimension of the concatenated feature maps to the same channel dimension as F_4 's, effectively lowering computational complexity (GFLOPs) and resource consumption. This step can be formally expressed as

$$F_{\text{all}} = W_1(\text{AAP}(\text{Concat}(F_1, F_2, F_3, F_4, F_5), C_4))(24)$$

where W_1 represents the weights of the 1×1 convolution, and C_4 denotes the target channel size used for channel unification.

The second step is ASS, which is essential for detecting ponding regions that exhibit a significant variation in spatial extents. Ponding may appear as small puddles or as large, extensive waterlogged areas, necessitating flexible scale adaptation. To address this variability, the ASS module employs convolution with receptive field sizes 3×3 , 5×5 , and 7×7 to generate three distinct feature maps. In addition, a sequential convolution operation using 3×3 , 5×5 , and 7×7 kernels produced a fourth feature map that aggregates contextual information across all scales. The final aggregated feature map F_{ASS} is computed as a weighted sum of these four feature maps, where the weights w_k are dynamically learned to emphasize the most relevant scales for accurate ponding detection:

$$F_{\text{ASS}} = \sum_{k \in \{k_1, k_2, k_3, \text{chain}\}} w_k \cdot F_k(X) (25)$$

where $F_{k_1}(X)$, $F_{k_2}(X)$, and $F_{k_3}(X)$ represent the feature maps generated by the 3×3 , 5×5 , and 7×7 convolution kernels, respectively, while $F_{\text{chain}}(X)$ denotes the feature map produced by their sequential application. The dynamically learned weights w_k enable the network to concentrate on the most relevant scales, thereby enhancing its capability to detect ponding regions of varying spatial sizes.

After the multi-scale features are aggregated through the ASS process, the next phase is PSR. This step is employed to refine these aggregated features. This stage incrementally enhances spatial accuracy and boundary clarity, a particularly important aspect in foggy conditions where blurred boundaries can obscure ponding regions. To address this challenge, PSR employs a series of depthwise separable convolutions to progressively refine the aggregated feature map F_{ASS} , improving spatial resolution while maintaining the structural consistency of multi-scale features.

During each refinement step, PSR divides the feature channels into two groups: a subset of distilled channels that undergo targeted refinement to capture fine spatial details, and the remaining channels, which reserve broader contextual

$$F_{i,\text{rem}}, F_{i,\text{dis}} = \begin{cases} \text{Split}_i(\mathcal{DL}_i(F_{\text{in}})) & i = 1 \\ \text{Split}_i(\mathcal{DL}_i(F_{\text{in}})) & i = 2, 3 \\ \mathcal{DL}_N(F_{N-1,\text{dis}}) & i = 4 \end{cases} (26)$$

where $F_{i,\text{rem}}$ represents the remaining channels at each refinement step, and $F_{i,\text{dis}}$ denotes the distilled channels that are progressively processed in subsequent layers. $\mathcal{DL}_i(\cdot)$ refers to the depthwise separable convolution with a leaky ReLU activation applied at the i th step, while $\text{Split}_i(\cdot)$ denotes the channel separation operation. N is the total number of iterative refinement steps.

After completing the final refinement stage, the distilled channels $F_{i,\text{dis}}$ obtained from all iterations are concatenated to construct the final distilled feature map:

$$F_{\text{distilled}} = \text{Concat}(F_{1,\text{dis}}, F_{2,\text{dis}}, F_{3,\text{dis}}, F_{4,\text{dis}}) (27)$$

The output $F_{\text{distilled}}$ consolidates the iterative enhancements, ensuring that fine spatial details and object boundaries are preserved with high clarity, even under challenging foggy conditions.

Beyond enhancing spatial details, the PSR module integrates a contrast-aware channel attention (CCA) mechanism. This mechanism enhances the model's capacity to prioritize channels based on their contrast sensitivity, a critical factor in foggy environments where the distinction between waterlogged and dry areas is often diminished. By focusing channels with higher contrast and suppressing less informative ones, CCA directs the network's focus toward the most discriminative features essential for accurate ponding detection. The structure of the CCA mechanisms is illustrated in Fig. 5(c).

The CCA mechanism determines attention weights by computing the global contrast for each channel. It captures both the mean and variance of the feature map to represent contrast variations effectively. These contrast descriptors are subsequently processed through learnable weight matrices, followed by non-linear activation functions to generate the attention map ζ , which selectively enhances channels with higher contrast:

$$\zeta = \sigma(W_3 \cdot (\phi(W_2 \cdot \text{Contrast}(F_{\text{distilled}})))) (28)$$

where $\text{Contrast}(F_{\text{distilled}})$ computes the mean and variance for each channel in $F_{\text{distilled}}$, thereby capturing the contrast profile. The resulting attention map ζ is subsequently applied to the feature map through element-wise multiplication:

$$F_{\text{CCA}} = \zeta \odot F_{\text{distilled}} (29)$$

where $\odot$ denotes element-wise multiplication.

By selectively amplifying channels that are sensitive to contrast variations, the CCA mechanism enables the model to focus on subtle yet crucial features necessary for accurate ponding detection, particularly under low-contrast and foggy conditions. This targeted attention enhances the model's robustness and accuracy, allowing it to effectively identify ponding areas even when visual degradation obscures essential scene details.

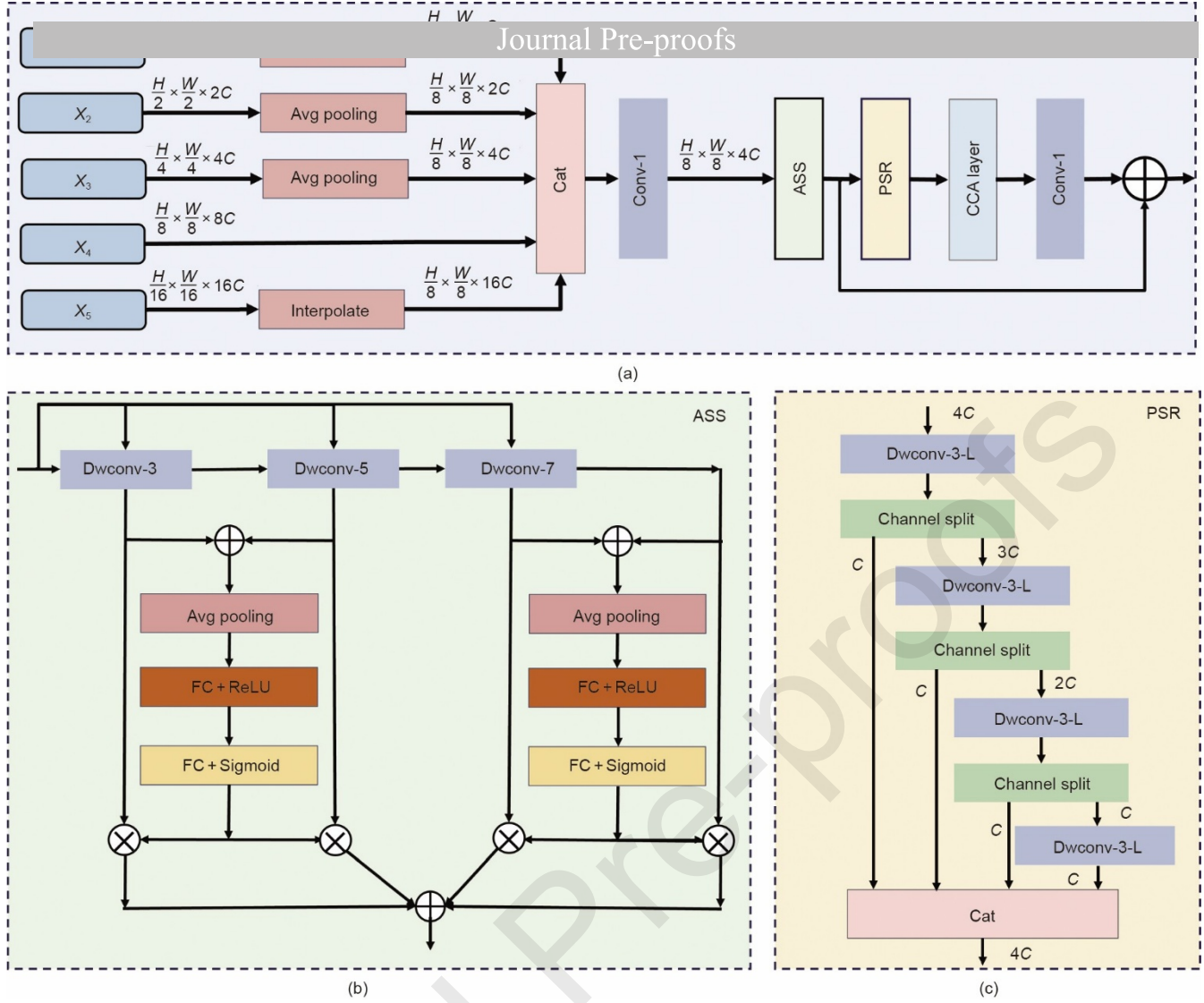


Fig. 6. Architecture of the MIA module. (a) Overall computational process of the MIA module. (b) Computational specifics of the ASS. (c) Progressive separable refinement module. Avg is the pooling stands for average pooling

3.4.2. AACG

The AACG module is developed to enhance feature fusion within the skip connections by adaptively regulating the information flow between the MIA module and each encoding stage. This selective regulation ensures that only contextually significant information is preserved and emphasized, which is essential for complex tasks, such as ponding detection under foggy environments, where specific spatial features need emphasis.

In the AACG module, feature fusion begins through a multi-head cross attention mechanism that integrates feature maps from the MIA module and the corresponding encoding stage. Let F_{MIA} and F_{encoder} represent the respective input feature maps. The query (Q), key (K), and value (V) matrices for each attention head i are computed as

$$\mathbb{Q}_i = W_Q^i F_{\text{MIA}}, \mathbb{K}_i = W_K^i F_{\text{encoder}}, \mathbb{V}_i = W_V^i F_{\text{encoder}} \quad (30)$$

where W_Q^i , W_K^i and W_V^i denote the learnable weight matrices for projecting the input features to the query, key, and value spaces of the i th attention head, respectively.

The multi-head self-attention output is subsequently obtained by concatenating the outputs of all attention heads and applying a final transformation:

$$\text{MHSA}(Q, K, V) = \text{Concat}(\text{head}_1, \text{head}_1, \dots, \text{head}_h) \quad (31)$$

where d_k represents the dimensionality of the key vector for each attention head. head_i denotes the i th attention head in the multi-head cross attention mechanism; $\text{Softmax}(\cdot)$ represents the softmax activation function; The final fused representation effectively integrates the most relevant contextual information from both the MIA module and encoding stage feature maps, capturing critical spatial and structural details.

To further refine this fused output, the AACG module applies layer normalization to F_{fused} , which stabilizes the feature distribution and enhances training robustness. F_{fused} denotes the fused feature map generated by the multi-head cross attention mechanism that integrates global contextual features from the MIA module and local spatial features from the corresponding encoder stage. This normalized feature map is then passed through a sigmoid activation function to generate a gating signal λ ranging from 0 to 1. This signal acts as an adaptive control gate, modulating the contribution of each feature channel according to its contextual importance. The sigmoid-activated gating signal is computed as

$$\lambda = \sigma(\text{LayerNorm}(F_{\text{fused}})) \quad (33)$$

where σ represents the sigmoid function, and $\text{LayerNorm}(\cdot)$ represents the layer normalization.

In the final stage of the AACG module, the generated gating signal is applied to the encoder feature map. Through element-wise multiplication, this gating signal selectively enhances or attenuates the features within F_{fused} , resulting in the final gated output:

$$F_{\text{gated}} = \lambda \odot F_{\text{encoder}} \quad (34)$$

To retain the original contextual information from the MIA module while integrating the adaptively gated features, a residual connection is established between the MIA feature map and the gated output. The final output of the AACG module, F_{output} , is thus computed by adding F_{MIA} back to F_{gated} :

$$F_{\text{output}} = F_{\text{MIA}} + F_{\text{gated}} \quad (35)$$

This residual connection preserves essential spatial information from the MIA module while enabling the network to adaptively emphasize contextually significant details. The gated output ensures that only the most relevant features are retained and propagated through the skip connection, thereby enhancing the network's capability to maintain critical spatial and contextual information.

In summary, the AACG module functions as an adaptive feature gate that integrates the most pertinent features from both the MIA module and the encoding stage feature maps through the combined use of cross attention, layer normalization, and sigmoid-based gating, followed by a residual connection. This fusion strategy allows the network to selectively focus on the most important features, ensuring robust and accurate ponding detection in complex environments, such as foggy environments, where the preservation of subtle features is vital for improving detection performance.

4. Experiments

In this section, we first introduce the datasets: Puddle-1000, Foggy-Puddle, and Foggy Low-Light Puddle. We then describe the experimental setup and the evaluation metrics used for performance assessment. Subsequently, a comparative analysis is carried out against SOTA methods on these datasets. To investigate the contribution of individual components to the overall network performance, ablation studies are performed. Finally, we evaluate the computational efficiency of the proposed network on an edge computing platform and provide a detailed analysis of the obtained results.

4.1. Datasets, settings, and evaluation metrics

4.1.1. Datasets

We evaluate the proposed method using three datasets: Puddle-1000, Foggy-Puddle, and Foggy Low-Light Puddle, to provide a comprehensive assessment of its effectiveness and robustness under challenging environmental conditions.

Puddle-1000: The Puddle-1000 dataset, introduced by Han et al. **Error! Reference source not found.**, is one of the few publicly available benchmarks specifically designed for road ponding detection in normal, clear-weather conditions. It consists of 985 manually annotated RGB images captured with a ZED camera (Stereolabs, USA), serving as a critical resource for evaluating algorithmic performance across both structured and unstructured environments. The dataset is divided into two subsets corresponding to different driving scenarios: the on-road subset, containing 357 structured ponding images, and the off-road subset, containing 628 unstructured images, both with a resolution of 640×360 . This division

Foggy-Puddle: Building upon the Puddle-1000 dataset, the Foggy-Puddle dataset was generated using an atmospheric scattering model, as described in Eq. (36), to simulate foggy conditions **Error! Reference source not found.** This dataset is constructed based on the atmospheric scattering equation:

$$I(x) = J(x) \cdot t(x) + A \cdot (1 - t(x)) \quad (36)$$

where $I(x)$ represents the foggy image, $J(x)$ is the clear-weather image, $t(x)$ denotes the medium transmission map, and A indicates the global atmospheric light. The transmission map $t(x)$ is generally computed using the following formula:

$$t(x) = e^{-\kappa d(x)} \quad (37)$$

where κ is the scattering coefficient that determines the fog density, and $d(x)$ represents the scene's depth at pixel location x . Eq. (37) models the exponential attenuation of light as it propagates through the fog, where higher values of κ correspond to denser fog and stronger light scattering. The Foggy-Puddle dataset is designed to serve as a challenging benchmark for evaluating road ponding detection under adverse weather conditions, particularly those involving fog-induced visibility degradation. By transforming the clear-weather images from the Puddle-1000 dataset into foggy conditions, it enables a comprehensive assessment of algorithm robustness against atmospheric distortions.

Foggy Low-Light Puddle: To extend the scope of road ponding detection to challenging low-light and reduced-visibility scenarios, we developed the Foggy Low-Light Puddle dataset as a dedicated benchmark for algorithm evaluation under such conditions. The dataset builds upon the night-pond dataset, which consists of 500 images captured under low-light and near-darkness conditions across urban, suburban, and campus roads in 10 cities in China.

We employed the Monodepth2 algorithm **Error! Reference source not found.** to generate accurate depth maps from the night-pond images, which were then used to guide the application of an atmospheric scattering model for simulating realistic fog effects. The depth information played a crucial role in this process, enabling adaptive adjustment of fog density according to object distance within the scene, producing denser fog in distant areas while maintaining details in closer areas. This depth-aware synthesis effectively captures the combined challenges of foggy and low-light conditions, establishing a robust foundation for evaluating algorithm performance under such compounded scenarios. By integrating both fog and low-light elements into a single benchmark, the Foggy Low-Light Puddle dataset fills a significant gap in road ponding research, providing a valuable resource for developing and validating algorithms in adverse environments. It serves as an important benchmark for advancing multi-condition road safety research, particularly in detecting ponding under foggy, low-light conditions.

4.1.2. Setting

Data preprocessing: In our experiments, all models were trained from scratch without the use of any pre-trained weights. During the preprocessing stage, all input images and their corresponding target masks were uniformly resized to 256×256 pixels. To ensure training stability, pixel intensity values were normalized to the range $[0, 1]$.

Training procedure: Training was initialized using the AdamW optimizer with an initial learning rate of 1×10^{-4} . A cosine annealing schedule was applied to progressively reduce the learning rate, enabling the model to escape local minima and achieve improved convergence. The batch size was set to 8, and training was conducted for 100 epochs or until early stopping was activated based on validation loss.

Computational environment: All experiments were implemented in PyTorch 2.1.1 (Meta, USA), and model evaluations were conducted on a single NVIDIA GeForce RTX 4090 GPU (NVIDIA, USA). The experimental environment was configured with CUDA 11.8, Python 3.10.13, and cuDNN 8.7.0.

4.1.3. Evaluation metrics

To quantitatively assess the performance of our segmentation model in road ponding detection, we used four widely recognized metrics: IoU, F1 score, mean IoU (MIoU), and mean pixel accuracy (MPA). These metrics collectively provided a comprehensive evaluation of the model's effectiveness and robustness.

IoU: IoU is a commonly used metric for assessing the degree of overlap between the predicted segmentation and the ground truth. It is defined as the ratio of the intersection (true positives) to the union of the predicted and ground truth regions (true positives, false positives, and false negatives). Mathematically, IoU is formulated as

$$\text{IoU} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}} \quad (38)$$

F1 score: The F1 score represents the harmonic mean of precision and recall, providing a balanced measure of accuracy between the two. It is especially valuable in cases of class imbalance, ensuring that both false positives and false negatives are effectively accounted for. The F1 score is defined as

$$F1 = \frac{2 \times TP}{2 \times TP + FP + FN} \quad (39)$$

MIoU: MIoU generalizes the IoU metric for multi-class segmentation by averaging the IoU values computed for each class. It provides a holistic measure of segmentation performance across all categories. It is calculated as

$$MIoU = \frac{1}{C_{cls}} \sum_{i=1}^C \frac{TP_i}{TP_i + FP_i + FN_i} \quad (40)$$

where C_{cls} represents the total number of classes, and TP_i , FP_i , and FN_i denote the true positives, false positives, and false negatives for each class i , respectively.

MPA: MPA evaluates the mean pixel-level accuracy across all classes. It accounts for both true positives and true negatives predictions, offering a detailed understanding of the model's performance in pixel-wise classification. The formula for MPA is

$$MPA = \frac{1}{C_{cls}} \sum_{i=1}^C \frac{TP_i + TN_i}{TP_i + TN_i + FP_i + FN_i} \quad (41)$$

where TN_i represents the true negatives for class i , while the remaining terms are defined as previously described. MPA ensures a balanced evaluation by accounting for both foreground (water) and background (non-water) regions, ensuring a comprehensive assessment of pixel-wise classification accuracy.

4.2. Comparison with SOTA methods

To assess the performance of ABCDWaveNet, we performed an extensive comparison with 11 SOTA methods, encompassing classical architectures, recent developments, and advanced approaches in road ponding detection and related domains. The compared methods included F3Net **Error! Reference source not found.**, CCNet **Error! Reference source not found.**, U2Net **Error! Reference source not found.**, SegFormer-B2 **Error! Reference source not found.**, SeaFormer-Base **Error! Reference source not found.**, UDTransNet **Error! Reference source not found.**, and VW-Former **Error! Reference source not found.** for general semantic segmentation; BWG **Error! Reference source not found.** for foggy semantic segmentation; and SWNet **Error! Reference source not found.**, Homo-Fusion **Error! Reference source not found.**, and AGSENet **Error! Reference source not found.** for road ponding detection. To maintain fairness, all experiments were conducted using publicly available data under a consistent environmental setup, and all prediction maps were evaluated using a unified codebase. The training procedure is shown in Fig. 7.

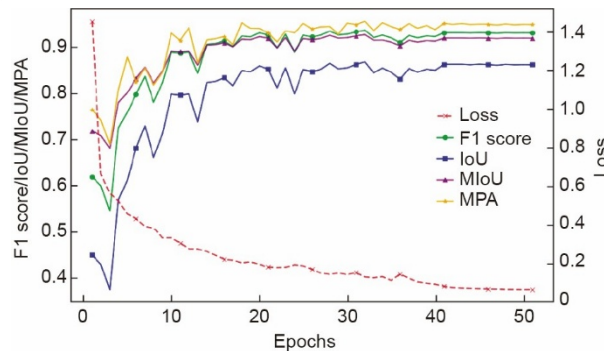


Fig. 7. Loss and performance metrics over training epochs on the Foggy-Puddle dataset.

4.2.1. Foggy-Puddle

Table 1 [31,32,34,49,58–64] presents a comprehensive comparison of various methods on the Foggy-Puddle dataset. The proposed ABCDWaveNet demonstrated SOTA performance, significantly outperforming existing methods. Specifically, ABCDWaveNet achieved an IoU of 84.59%, surpassing the second-best method, BWG **Error! Reference source not found.**, by 3.51% points, and a F1 score of 92.22%, exceeding BWG by 2.98% points. The qualitative comparison in Fig. 8

Beyond its accuracy, ABCDWaveNet also delivers competitive efficiency, with an inference speed (FPS) of 25.48 on the NVIDIA Jetson AGX Orin (NVIDIA, USA). This combination of high detection accuracy and computational efficiency establishes ABCDWaveNet as a practical and effective solution for ADAS early-warning systems operating in adverse weather conditions.

4.2.2. Puddle-1000

Table 2 [31,32,34,49,58–64] presents the performance results on the Puddle-1000 dataset, which evaluates road ponding segmentation under normal weather conditions. The proposed ABCDWaveNet achieved the highest IoU of 85.84%, exceeding AGSENet **Error! Reference source not found.**, the second-best performance method, by 1.75% points. Similarly, ABCDWaveNet attained an F1 score of 93.06%, improving upon AGSENet by 1.70% points. Moreover, it recorded an MIoU of 92.45% and an MPA of 96.89%, reflecting respective gains of 0.49% and 1.57% points.

4.2.3. Foggy Low-Light Puddle

To further evaluate the robustness of ABCDWaveNet under compounded adverse conditions, the model was tested on the Foggy Low-Light Puddle dataset (Table 3). This dataset integrates both foggy and low-light conditions, posing substantial challenges for segmentation accuracy. ABCDWaveNet achieved an IoU of 62.51%, surpassing BWG **Error! Reference source not found.** by 1.03 percentage points, demonstrating its strong generalization and adaptability in visually degraded environments. The F1 score achieved was 78.84%, which was 1.01 percentage points higher than that of BWG. The MIoU and MPA were 76.93% and 88.66%, respectively, demonstrating consistent improvements over existing methods.

Experimental results across all datasets confirmed that ABCDWaveNet consistently outperformed existing SOTA methods in road ponding segmentation, particularly under adverse weather conditions. The significant improvements in IoU and F1 score indicated that the proposed model effectively captured and differentiated ponding features, even when visibility was reduced. Although ABCDWaveNet exhibits a moderate parameter count and computational cost, its substantial gains in segmentation accuracy justify these requirements, making it a strong candidate for safety-critical applications, such as ADASs.

Table 1

Comprehensive quantitative comparison of SOTA methods on the Foggy-Puddle dataset.

Methods	Publication year	Performance				#Param (M)	GFLOPs	FPS
		IoU (%)	F1 score (%)	MIoU (%)	MPA (%)			
FCMNet Error!	2020	77.42	88.65	88.64	94.57	75.54	8.48	21.07
CCNet Error!	2020	79.55	89.15	89.78	95.20	71.27	79.69	22.33
U2Net Error!	2020	80.20	90.10	90.07	94.06	1.77	12.07	25.51
SeeFormer Base Error!	2021	78.26	88.47	88.17	94.96	27.45	14.00	24.71
SWNet Error!	2022	70.44	88.02	88.99	95.14	21.91	26.94	28.92
Homofusion Error!	2022	78.25	88.65	88.72	94.92	1.24	61.20	25.41
SeeFormer Base Error!	2022	78.06	88.61	88.29	95.22	9.59	8.42	42.69
UDTTransNet Error!	2024	76.56	88.11	86.72	94.50	22.02	24.59	19.64
UWFFormer Error!	2024	76.72	88.16	86.92	93.79	84.74	27.20	19.60
AGSENet Error!	2024	78.20	88.55	88.45	93.22	2.05	16.01	22.74
BWG Error! Reference	2024	81.09	90.24	90.25	95.15	72.46	82.54	15.52

Metrics shown in ABCDWaveNet (Ours) bold represent the best performance. We assessed the performance of the model using multiple evaluation metrics, including IoU, F1 score, MIoU, and MPA. In addition, we report the parameter count (#Param in millions), GFLOPs, and FPS. The FPS was measured on an NVIDIA Jetson AGX Orin with an input image resolution of 256×256 .

Table 2
Comprehensive quantitative comparison of SOTA methods on the Puddle-1000 dataset.

Methods	Publication year	Puddle-1000				#Param (M)	GFLOPs	FPS
		IoU (%)	F1 score (%)	MIoU (%)	MPA (%)			
E2Net E2Net†	2020	81.84	90.01	90.70	94.36	25.54	0.48	21.07
CCNet CCNet†	2020	82.22	91.02	90.72	94.40	71.27	70.60	22.22
U2Net U2Net†	2020	82.57	91.22	90.69	95.17	1.77	12.07	25.51
SeaFormer Base SeaFormer†	2021	90.40	90.10	90.12	94.55	27.45	14.00	24.71
SWNet SWNet†	2022	92.01	90.11	90.90	94.77	21.91	26.94	29.92
HomeFusion HomeFusion†	2022	90.26	90.12	90.01	95.67	1.24	61.20	25.41
SeaFormer Base SeaFormer†	2022	91.22	90.10	90.25	94.00	9.59	0.42	42.69
U2TetraNet U2TetraNet†	2024	77.50	99.64	97.29	95.56	22.02	24.59	19.64
VWFoformer VWFoformer†	2024	77.09	90.27	99.70	94.64	94.74	27.20	10.60
DW3 DW3† <i>Reference</i>	2024	92.04	91.27	91.25	95.22	72.46	92.54	15.52
AGSENet AGSENet†	2024	94.00	91.26	91.06	95.22	2.05	16.01	22.74
ABCDWaveNet (Ours)	2026	85.84	93.06	92.45	96.89	47.51	38.16	25.48

Metrics shown in ABCDWaveNet bold denote the best performance. We evaluated the model using multiple metrics, including IoU, F1 score, MIoU, and MPA, along with the parameter count (#Param, in millions), GFLOPs, and FPS. The FPS was measured on an NVIDIA Jetson AGX Orin with an input image resolution of 256×256 .

Table 3
Comprehensive quantitative comparison of SOTA methods on the Foggy Low-Light Puddle dataset.

Methods	Publication year	Foggy Low-Light Puddle				#Param (M)	GFLOPs	FPS
		IoU (%)	F1 score (%)	MIoU (%)	MPA (%)			
E2Net E2Net†	2020	55.44	72.40	72.69	91.01	25.54	0.48	21.07
CCNet CCNet†	2020	57.74	72.19	72.24	92.29	71.27	70.60	22.22
U2Net U2Net†	2020	59.62	74.70	77.15	92.59	1.77	12.07	25.51
SeaFormer Base SeaFormer†	2021	55.10	71.05	74.16	92.77	27.45	14.00	24.71
SWNet SWNet†	2022	51.20	67.90	71.70	92.55	21.91	26.94	29.92
HomeFusion HomeFusion†	2022	59.62	75.45	74.02	92.09	1.24	61.20	25.41
SeaFormer Base SeaFormer†	2022	60.50	75.20	77.40	95.67	9.59	0.42	42.69
U2TetraNet U2TetraNet†	2024	60.00	77.14	75.07	97.56	22.02	24.59	19.64
VWFoformer VWFoformer†	2024	60.06	77.57	76.16	97.09	94.74	27.20	10.60
AGSENet AGSENet†	2024	50.11	74.60	76.91	92.10	2.05	16.01	22.74
DW3 DW3† <i>Reference</i>	2024	61.49	77.92	76.22	99.22	72.46	92.54	15.52
ABCDWaveNet (Ours)	2026	62.51	78.84	76.93	88.66	47.51	38.16	25.48



Fig. 8. Representative prediction results on the Foggy-Puddle dataset, comparing ABCDWaveNet (ours), AGSENet, BWG, SWNet, Homofusion, and U2Net. The bottom row shows the ground-truth annotations. Yellow boxes denote missed detections, while blue boxes indicate false detections.

4.3. Ablation studies

In our comprehensive ablation experiments on the Foggy-Puddle dataset, we investigated the individual and combined effects of three core components in the proposed model: the BIS module, the MIA module, and the AACG module. The quantitative results summarized in Table 4 demonstrate the contribution of each component to puddle detection performance under foggy conditions. Furthermore, we first investigate the effect of the layer architecture on model performance in Table 5, followed by an analysis of the impact of varying the number of dynamic experts in the DDC layer on the overall performance and efficacy of ABCDWaveNet in Table 6. We then present a detailed evaluation of the different convolution types within the DDC layer in Table 7, and finally assess the performance of distinct wavelet basis function types for the frequency-aware pathway in Table 8.

Table 4
Ablation study on the proposed components of ABCDWaveNet.

Models	IoU (%)	F1 score (%)	MIoU(%)	MPA(%)	#Param (M)	GFLOPs
ABCDWaveNet (full model)	84.59	92.22	91.65	95.74	47.51	38.16
w/o BIS	83.72 (−0.87)*	91.78 (−0.44)*	91.14 (−0.51)*	95.34 (−0.40)*	37.33 (−10.18)#	28.62 (−9.54)#
w/o ASS	84.16 (−0.43)*	92.01 (−0.21)*	91.40 (−0.25)*	95.65 (−0.09)*	47.50 (−0.01)#	37.71 (−0.45)#
w/o PSR	84.25 (−0.34)*	92.05 (−0.17)*	91.47 (−0.18)*	95.70 (−0.04)*	47.24 (−0.27)#	37.90 (−0.26)#
w/o AACG	84.04 (−0.55)*	91.99 (−0.23)*	91.37 (−0.28)*	95.59 (−0.15)*	47.01 (−0.50)#	26.32 (−11.84)#
w/o BIS + AACG	83.67 (−0.92)*	91.76 (−0.46)*	91.11 (−0.54)*	95.30 (−0.44)*	36.84 (−10.67)#	16.78 (−21.38)#
w/o MIA + AACG	83.86 (−0.73)*	91.86 (−0.36)*	91.23 (−0.42)*	95.36 (−0.38)*	45.69 (−1.82)#	25.01 (−13.15)#

* values indicate a performance deterioration relative to the baseline configuration, whereas # values represent improvements. Bold values highlight the best performance, and w/o stands for without. We evaluated the IoU, F1 score, MIoU, and MPA metrics on the Foggy-Puddle dataset while also measuring the parameter size and GFLOPs. PSR, Progressive Separable Refinement.

Table 5

Ablation study on different configurations of layers in a dynamic convolution layer.

Model architecture	IoU(%)	F1 score(%)	MIoU(%)	MPA(%)	#Param (M)	GFLOPs
Dual-layer	84.59	92.22	91.65	95.74	47.51	38.16
Single-layer	81.13 (−3.46)*	89.06 (−3.16)*	90.38 (−1.27)*	93.80 (−1.94)*	27.07 (−20.44)#	21.77 (−16.39)#

* values indicate a degradation relative to the baseline configuration, while # values denote an improvement. Bold values represent the best performance.

Table 6

Ablation study on different configurations of experts in dynamic convolution.

Num_experts	IoU(%)	F1 score(%)	MIoU(%)	MPA(%)	#Param (M)	GFLOPs
[2, 2, 2, 4, 4]	84.59	92.22	91.65	95.74	47.51	38.16
[2, 2, 2, 2, 2]	84.05 (−0.54)*	91.95 (−0.27)*	91.33 (−0.32)*	95.55 (−0.09)*	30.99 (−16.52)#	38.14 (−0.02)#
[4, 4, 4, 2, 2]	84.21 (−0.38)*	92.04 (−0.18)*	91.44 (−0.21)*	95.66 (+0.02)#	33.28 (−16.52)#	38.14 (−0.02)#
[4, 4, 4, 4, 4]	84.62 (+0.03)#	92.28 (+0.06)#	91.70 (+0.05)#	95.68 (+0.04)#	49.80 (+2.29)*	38.16 (+0.00)#

* values indicate a degradation relative to the baseline configuration, while # values denote an improvement. Bold values highlight the best performance.

Table 7

Ablation study on different convolution types in dual convolutional layers.

Convolution type	IoU(%)	F1 score(%)	MIoU(%)	MPA(%)	#Param (M)	GFLOPs
Dynamic convolution	84.59	92.22	91.65	95.74	47.51	38.16
Standard convolution	82.33 (−2.26)*	90.94 (−1.28)*	89.93 (−1.72)*	95.16 (−0.58)*	21.58 (−25.93)#	38.11 (−0.05)#
Dilated convolution	79.49 (−5.10)*	89.64 (−2.58)*	88.57 (−3.08)*	94.09 (−1.65)*	21.58 (−25.93)#	38.11 (−0.05)#
Group convolution	79.95 (−4.64)*	89.88 (−2.34)*	88.86 (−2.79)*	95.08 (−0.66)*	14.56 (−32.95)#	29.05 (−9.11)#

* values indicate a deterioration relative to the baseline configuration, while # values indicate an improvement. Bold values represent the best performance.

Table 8

Ablation study on different types of wavelet basis functions.

Wavelet basis function	IoU(%)	F1 score(%)	MIoU(%)	MPA(%)	FPS
Haar	84.59	92.22	91.65	95.74	25.48
Daubechies (db2)	84.17 (−0.42)*	91.75 (−0.47)*	92.43 (+0.78)#	95.20 (−0.54)*	14.07 (−11.41)*
Symlet (sym2)	84.48 (−0.11)*	92.39 (+0.17)#	92.12 (−0.53)*	95.54 (+0.30)#	9.86 (−15.62)*
Coiflet (coif2)	84.79 (+0.20)#	92.84 (+0.62)#	92.39 (+0.74)#	95.87 (+0.13)#	5.19 (−20.29)*

* values indicate a deterioration relative to the baseline configuration, while # values indicate an improvement. Bold values represent the highest performance.

4.3.1. Analysis of BIS module

The BIS module plays a vital role in improving the model’s robustness and accuracy by simultaneously processing low- and high-frequency components. It integrates a frequency-aware information pathway that focuses on low-frequency components to preserve the global structural consistency of the image, while high-frequency features enhance fine details and capture complex, intricate textures. In addition, the spatial-enhanced feature learning pathway strengthens the model’s ability to localize objects and interpret spatial relationships across the scene. This function is particularly crucial in visually complex conditions where subtle variations in texture and structure must be detected. Through these dual processing mechanisms, the BIS module ensures the preservation of global structural information while enhancing the fine-grained details, which are critical for precise road ponding detection.

To evaluate the BIS module’s impact, we performed an ablation study using a model variant without the BIS component module (denoted as “w/o BIS,” as presented in Table 4). The removal of the BIS module led to a notable reduction in performance. The IoU decreased from 84.59% to 83.72%, a reduction of 0.87%, and the F1 score declined from 92.22% to 91.78%, a loss of 0.44%. These findings clearly indicate the BIS module’s significant contribution, particularly in retaining detailed structural and fine-grained details in the presence of environmental disturbances, such as haze or fog.

In addition to the observed performance decline, the removal of the BIS module resulted in a reduction of the model’s parameter count from 47.51 million to 37.33 million and a decrease in GFLOPs from 38.16 to 28.62. These changes indicated improved computational efficiency; however, this efficiency was achieved at the expense of a notable drop in accuracy. Although the reduction in GFLOP may appear advantageous, the loss in segmentation performance emphasizes the indispensable role of the BIS module. Its inclusion is well justified, as the module’s contribution to maintaining high detection accuracy in challenging environments significantly outweighs the benefits of lower computational cost.

The heatmaps (Fig. 9, row 2) further substantiated these observations, showing that, in the absence of BIS, the model failed to preserve strong activation in key regions, such as puddle edges and water reflections, where high-frequency features are critical for precise boundary delineation. In contrast, the model incorporating the BIS module exhibited enhanced activation in these regions, highlighting the model’s capability to preserve and enhance essential fine-grained features. This visual evidence, combined with the quantitative performance results, underscores the crucial role of the BIS module in ensuring the overall effectiveness and robustness of the model.

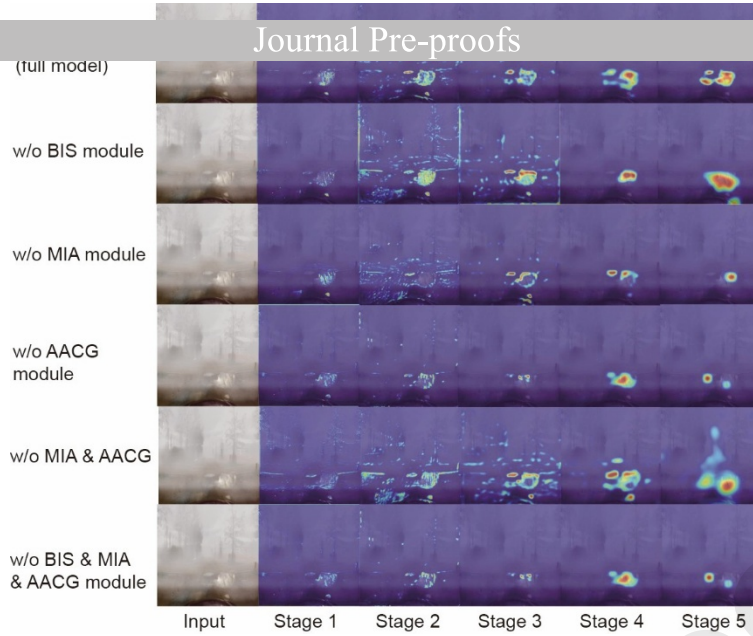


Fig. 9. Heatmaps of encoder activations across stages (Stages 1–5), arranged from left to right and top to bottom for the following configurations: ABCDWaveNet (full model), w/o BIS module, w/o MIA module, w/o AACG module, w/o MIA & AACG, and w/o BIS & MIA & AACG module.

4.3.2. Analysis of ABC mechanism

Our detailed analysis of the ABC mechanism highlighted its well-structured design for effectively integrating multi-scale information. The ABC mechanism comprises two essential components: the MIA and AACG modules. The MIA module is specifically designed to combine features from different encoding stages in a coarse-to-fine manner, allowing the model to generate rich and detailed representations through efficient multi-scale feature fusion. This strategy enables the model to capture both broad structural patterns and fine-grained details, which are crucial for accurate detection in visually complex environments.

To evaluate the contribution of the MIA module in progressively capturing multi-scale information, an ablation study was conducted by removing both the ASS and PSR modules. The ASS module plays a key role in learning structural and contextual information at a coarse-grained level, whereas the PSR module extracts highly discriminative features to refine spatial accuracy. The results of this study were reflected in the performance metrics, where removing the ASS module resulted in a decrease in IoU by 0.43%, while excluding the PSR module resulted in an additional drop of 0.34%. These reductions emphasize the significant contributions of both modules in maintaining multi-scale feature integrity and improving overall model accuracy.

The heatmap analysis further supported these results (Fig. 9, row 3), revealing diminished activation in critical regions, where multi-scale information is essential, particularly around water boundaries and subtle textural variations. The absence of the MIA module led to a less detailed representation, particularly in regions requiring fine-grained differentiation. The heatmap demonstrated a visible loss of localized detail when the multi-scale fusion process was disrupted, underscoring the importance of the MIA module.

We further investigated the role of the AACG module, which refines multi-scale features by adaptively weighting the importance of features from different stages. To assess its influence, the AACG module was removed (Fig. 9, row 4), and the concatenation operation was replaced with a simpler element-wise addition fusion method. As presented in Table 4, the complete model attained an IoU of 84.59%. However, when the AACG module was removed (denoted as “w/o AACG”), the model’s performance decreased noticeably, with the IoU dropping to 84.04%, a reduction of 0.55%. This result emphasizes the essential role of the AACG module in enhancing feature fusion and enabling the model to adaptively emphasize the most relevant information across different stages.

Subsequently, we examined the impact of removing both the MIA and AACG modules (denoted as “w/o MIA + AACG”). This configuration led to a more pronounced performance drop, with the IoU dropping to 83.86%, corresponding to a reduction of 0.73%. The heatmaps visualizations (Fig. 9, row 5) further corroborated these findings, showing weaker activation in essential regions and indicating a diminished ability of the model to effectively integrate and prioritize multi-scale information.

Overall, these results demonstrate that both the MIA and AACG modules are vital to the success of the ABC mechanism.

Their exclusion resulted in significant degradation in model accuracy and representational quality. The heatmap evidence further Journal Pre-proofs when these modules were omitted.

4.3.3. Analysis of layer architectures in dynamic convolutional networks

To evaluate the effectiveness of different layer architectures, we conducted an extensive ablation study, the results of which are summarized in Table 5. The findings indicated that the dual-layer architecture consistently and significantly surpassed the single-layer configuration across major performance metrics, achieving improvements of +3.46% in IoU and +3.16% in F1 score.

In our proposed design, each encoder block incorporates DDC layers. In contrast to the single-layer configuration, which employs a single convolution followed by batch normalization and activation, the dual-layer structure enables a more comprehensive feature extraction process. The first layer captures coarse, adaptive representations from the input, while the second layer refines these features through deeper extraction and an additional linear transformation. This process effectively broadens the receptive field and enhances fine-detail resolution.

Performing this dual-stage feature extraction before downsampling is essential for preserving both fine textures and high-level semantic information, which are often diminished during resolution reduction. The refinement is particularly advantageous under challenging visual conditions, such as fog, where scene visibility and edge detail are substantially degraded.

Although the dual-layer architecture introduces additional computational overhead, specifically, an increase in parameters from 27.07 M to 47.51 M and in GFLOPs from 21.77 to 38.16, these costs are justified by the considerable performance improvements achieved. The enhanced capability of the model to capture multi-scale and adaptive refined features results in superior accuracy for complex tasks, such as road ponding detection. This makes the dual-layer architecture particularly beneficial for safety-critical applications, such as ADAS, where precise and reliable detection is essential for ensuring operational safety.

4.3.4. Analysis of different configurations of dynamic experts

The configuration of dynamic experts plays a critical role in defining the overall capacity and GFLOP of the model. We performed an extensive evaluation of multiple expert configurations across the dual convolution layers, with the results summarized in Table 6. The analysis revealed that assigning two experts to the first three encoding stages and four experts to the final two stages (i.e., [2, 2, 2, 4, 4]) provided the most effective performance. This configuration strikes a delicate balance between model accuracy and computational efficiency, effectively utilizing the model’s capability to capture both low- and high-level features without adding unnecessary complexity.

In contrast, adopting a uniform configuration of only two experts across all stages, although computationally efficient, resulted in a notable decline in performance. The reduced representational capacity in the deeper layers, which are critical for extracting more abstract and complex features, led to a considerable decline in model accuracy. This finding indicates that a uniform and shallow expert configuration is not sufficient for addressing the complexities of high-level feature extraction, especially in tasks that demand fine-grained discriminative precision.

Furthermore, we examined a configuration with four experts assigned to the initial encoding stages and two experts to the deeper stages (i.e., [4, 4, 4, 2, 2]). This configuration resulted in a 0.38% decrease in IoU, suggesting that the deeper layers, which handle lower-resolution yet semantically richer features, require a larger number of experts to accurately capture and represent these features. Reducing the number of experts in these layers significantly weakened the model’s ability to encode complex, high-level information. This finding highlights the necessity of distributing experts in accordance with the semantic depth and feature resolution of each network stage.

Additionally, we investigated the effect of uniformly increasing the number of experts across all stages using the configuration [4, 4, 4, 4, 4]. Although this setup enhanced the model’s capacity to capture a broader range of features, it did not lead to a significant improvement in accuracy. This finding reveals an important insight: increasing the number of experts alone does not necessarily result in better performance. Instead, the added experts increase the overall model complexity, making the joint optimization of convolution kernels and attention mechanisms more difficult. This elevated complexity also raises the likelihood of overfitting, especially in cases where the model may capture irrelevant patterns or assign excessive importance to less informative features. Furthermore, adding more experts substantially raises the computational demands in terms of both memory usage and processing time. In resource-limited scenarios, such as real-time inference on embedded platforms, this higher computational cost may not be justified by the marginal gains in accuracy. Therefore, although a larger number of experts can theoretically improve representational power by allowing the model to learn a wider variety of features, it is essential to maintain a careful balance between computational efficiency and the risk for overfitting.

In summary, the configuration [2, 2, 2, 4, 4] achieves an optimal balance between model complexity, representational capacity, and computational cost. This configuration represents a well-calibrated design decision, supported by both theoretical reasoning and empirical validation, effectively satisfying performance requirements while preserving computational efficiency.

4.3.5. Analysis of convolution type in dual convolution layers

In this part of our ablation study, we evaluate the effect of different convolution types employed in the DDC layer on the performance of ABCDWaveNet. Four convolution types were compared: standard convolution, dilated convolution with a dilation rate of 2, group convolution with a group size of 4, and depthwise convolution. The corresponding results are presented in Table 7. Among these, dynamic convolution achieved the best performance, with an IoU of 84.59% and F1 score of 92.22%. Notably, the GFLOPs of dynamic convolution were nearly comparable to those of standard and dilated convolutions, exceeding them by only 2.26% and 5.10%, respectively. Although dilated convolution expands the receptive field, it compromises the ability to capture fine spatial details, resulting in reduced performance in scenarios that require complex feature extraction.

Group convolution demonstrated relatively better computational efficiency; however, this advantage came at the cost of decreased performance. Specifically, the IoU and F1 score decreased by 4.64% and 2.34%, respectively. The reduced parameterization limited feature integration across channels, hindering the model’s ability to capture intricate spatial relationships essential for high-precision tasks. Depthwise convolution yielded the lowest accuracy, with an IoU of 73.64% and F1 score of 86.68%. Although it offers the smallest parameter count and GFLOPs, its accuracy remains inadequate for high-precision applications, thereby restricting its effectiveness in complex environments that demand detailed feature extraction. Overall, dynamic convolution emerged as the most effective choice, delivering superior performance without a substantial increase in GFLOPs. It consistently maintained high accuracy in challenging conditions, such as foggy environments, where precise feature extraction is critical. In comparison, the other convolution types struggled to sustain comparable performances, emphasizing the distinct advantages of dynamic convolution in complex detection scenarios.

4.3.6. Analysis of wavelet basis function in frequency-aware pathway

In the experiments, we evaluated four widely used wavelet basis functions: Haar, Daubechies (db2 variant), Symlet (sym2 variant), and Coiflet (coif2 variant). The experimental outcomes for each wavelet basis function are summarized in Table 8. The Coiflet wavelet showed slight improvements in IoU and F1 score, with increases of 0.20% and 0.62%, respectively, compared with those of Haar. However, its FPS (5.19) was considerably lower than that of the other wavelets.

Given the stringent real-time requirements of ADAS early-warning systems, these minor accuracy gains are insufficient to compensate for the notable decrease in computational efficiency. Similarly, both Daubechies (db2) and Symlet (sym2) exhibited marginal improvements in certain metrics, such as MIOU for Daubechies and F1 score for Symlet, but their overall performance did not surpass that of Haar. Furthermore, their FPS values (14.07 and 9.86) were significantly lower than Haar’s, rendering them unsuitable for real-time applications, especially under challenging conditions, such as fog. In contrast, the Haar wavelet achieved the most favorable trade-off between accuracy and computational efficiency, obtaining an IoU of 84.59%, F1 score of 92.22%, and FPS of 25.48. This balance allows Haar to effectively meet the requirements of road ponding detection while maintaining timely response capabilities in complex environments. Therefore, the Haar wavelet basis function was selected as the optimal choice for this task, as it provides the best balance between accuracy and processing speed, which is essential for reliable performance in early-warning systems.

4.4. Analysis of model generalization

This section presents a detailed analysis of the generalization capability of ABCDWaveNet across varying fog densities, driving conditions, and road types. The model’s performance was examined under slight, medium, heavy, and low-light fog conditions and evaluated on both structured (e.g., urban streets, campus roads, and expressways) and unstructured roads to assess its robustness and potential limitations.

To evaluate cross-domain generalization, the weights trained on the Foggy-Puddle dataset (<https://github.com/Lyu-Dakang/ABCDWaveNet>) under dense fog conditions were used to test the model on the slight and medium fog datasets. This evaluation highlights the model’s ability to generalize effectively across fog densities, with the results summarized in Table 9. For the low-light fog condition, visual results were generated using weights trained on the Foggy Low-Light Puddle dataset, as illustrated in Fig. 10.

Table 9

Performance of ABCDWaveNet under different fog densities and conditions. The results for slight fog and medium fog were obtained by directly testing the model using the pre-trained weights trained on heavy fog conditions from the Foggy-Puddle dataset.

Slight fog	85.47	92.87	92.01	95.98	Foggy-Puddle	Heavy fog
Medium fog	85.16	92.38	91.48	95.77	Foggy-Puddle	Heavy fog
Heavy fog	84.59	92.22	91.65	95.74	Foggy-Puddle	Heavy fog
Low-light & slight fog	62.51	78.84	76.93	88.66	Foggy Low-Light Puddle	Low-light & slight fog



Fig. 10. ABCDWaveNet detection results across varying fog densities and road types, including slight fog, medium fog, heavy fog, and low-light & slight fog, demonstrating the model's robustness and generalization capability under diverse environmental conditions.

The findings indicated that ABCDWaveNet achieved excellent performance in puddle detection under slight and medium fog conditions, maintaining high IoU and F1 scores, especially in clear, unobstructed environments. As fog density increased, however, a moderate reduction in performance was observed. While the model consistently identified larger puddles, smaller puddles with low contrast or those partially occluded were occasionally missed, resulting in false negatives. Under low-light fog conditions, reflections and background noise can also contribute to false detections.

Despite these challenges, the model demonstrated strong generalization across a wide range of environmental conditions, as confirmed by both quantitative and qualitative analyses. Failure cases primarily involved occlusions caused by shadows or reflections and fragmented puddle patterns, which sometimes led to misdetections, particularly in campus road and expressway scenarios. In addition, in heavy fog conditions, reduced sensitivity to very small or thin water formations occurred due to frequency attenuation effects.

In conclusion, ABCDWaveNet exhibited strong generalization capabilities across different fog densities and a variety of road types, although there remains room for further improvements. Future research will aim to tackle the difficulties associated with detecting small targets and mitigating environmental interference, with possible strategies including the application of multi-scale techniques and the incorporation of infrared light information.

4.5. Model deployment

To further validate our algorithm in real-world conditions, we deployed and tested the model within an automotive setting, complementing evaluations on existing datasets. The vehicle-mounted edge computing platform, illustrated in Fig. 11, comprised an NVIDIA Jetson AGX Orin, a display monitor, and a data cable. As reported in Table 1, the model achieved a frame rate of 25.48 FPS. In addition, the model loading time was 2.62 s, and the inference latency exhibited a standard deviation of 3.79 ms over 200 runs, indicating high runtime stability and consistency. Peak GPU memory usage during inference reached 601.5 MB. These results demonstrate that the proposed method reliably delivered both high accuracy and operational efficiency on the vehicle's edge platform, effectively satisfying the real-time performance requirements of ADAS early-warning systems under constrained computational resources.

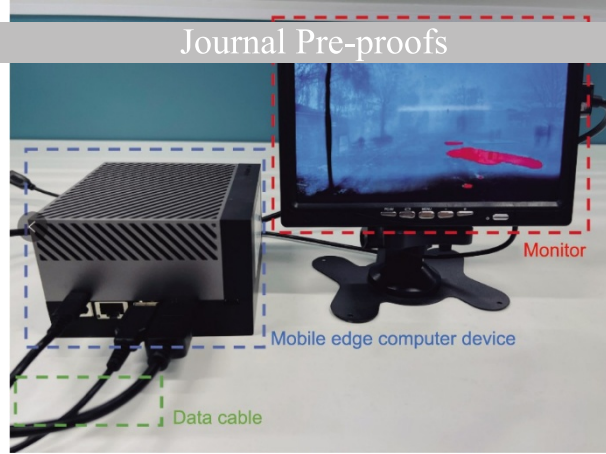


Fig. 11. An edge computing system built with an NVIDIA Jetson AGX Orin.

5. Conclusion

In this study, we introduced ABCDWaveNet, an advanced framework designed for robust road ponding detection under foggy conditions, utilizing the synergistic integration of dynamic convolution and wavelet-based frequency-spatial enhancement. Through the BIS module and ABC mechanism, the model effectively captures both global structure and fine-grained details across multiple scales. Comprehensive experiments conducted on the Puddle-1000, Foggy-Puddle, and newly proposed Foggy Low-Light Puddle dataset demonstrated that ABCDWaveNet achieved SOTA performance on edge devices, confirming its practicality for ADAS applications. It is anticipated that this study will provide meaningful insights for advancing resilient and proactive road safety systems under visually degraded conditions. To facilitate future research and ensure reproducibility, the code and datasets will be made publicly accessible to the research community.

Supplement Material

Our code is available publicly at <https://github.com/Lyu-Dakang/ABCDWaveNet>.

Acknowledgement

This project was jointly supported by the Open Fund Project of State Key Laboratory of Intelligent Green Vehicle and Mobility (KFZ260401), the National Natural Science Foundation of China (52172350, W2421069, and 51775565), the Guangzhou Science and Technology Plan Project (2024B01W0079), the Science and Technology Planning Project of Guangdong Province (2023B1212060029), and the China Postdoctoral Science Foundation (2013T60904).

Compliance with ethics guidelines

Ronghui Zhang, Dakang Lyu, Tengfei Li, Yunfan Wu, Ujjal MANANDHAR, Benfei Wang, Junzhou Chen, Bolin Gao, Danwei Wang, and Yiqiu Tan declare that they have no conflict of interest or financial conflicts to disclose.

References

- [1] Ali F, Khan ZH, Khattak KS, Gulliver TA. The effect of visibility on road traffic during foggy weather conditions. *IET Intell Transp Syst* 2024;18(1):47–57.
- [2] Wang H, Shao W, Sun C, Yang K, Cao D, Li J. A survey on an emerging safety challenge for autonomous vehicles: safety of the intended functionality. *Engineering* 2024;33:17–34.
- [3] Liu J, Xiong W, Bai L, Xia Y, Huang T, Ouyang W, et al. Deep instance segmentation with automotive radar detection points. *IEEE Trans Intell Veh* 2023;8(1):84–94.
- [4] Yan B, Fang C, Qiu H, Zhu W. Intelligent speed limit system for safe expressway driving in rainy and foggy weather based on internet of things. *J Shanghai Jiaotong Univ (Sci)* 2023;28(1):10–9.
- [5] Hao H, Wang Y, Chen J. Empowering scenario planning with artificial intelligence: a perspective on building smart and resilient cities. *Engineering* 2024;43:272–83.
- [6] U.S. Department of Transportation. Road weather management program. Report. Washington, DC: U.S. Department of Transportation; 2023.
- [7] Gong B, Wang F, Lin C, Wu D. Modeling HDV and CAV mixed traffic flow on a foggy two-lane highway with cellular automata and game theory model. *Sustainability* 2022;14(10):5899.

[9] Li J, Shao W, Wang H. Key challenges and chinese solutions for SOTIF in intelligent connected vehicles. *Engineering* 2023;31:27–30.

[10] Yan Y, Ni L, Sun L, Wang Y, Zhou J. Digital twin enabling technologies for advancing road engineering and lifecycle applications. *Engineering* 2025;44:184–206.

[11] Ma Y, Wang M, Feng Q, He Z, Tian M. Current non-contact road surface condition detection schemes and technical challenges. *Sensors* 2022;22(24):9583.

[12] Xiong W, Zou Z, Zhao Q, He F, Zhu B. Lxlv2: enhanced lidar excluded lean 3D object detection with fusion of 4d radar and camera. *IEEE Robot Autom Lett* 2025;10(3):2862–9.

[13] Xiong W, Liu J, Huang T, Han QL, Xia Y, Zhu B. Lxl: lidar excluded lean 3D object detection with 4D imaging radar and camera fusion. *IEEE Trans Intell Veh* 2024;9(1):79–92.

[14] Zheng L, Li S, Tan B, Yang L, Chen S, Huang L, et al. Rcfusion: fusing 4-D radar and camera with bird's-eye view features for 3-D object detection. *IEEE Trans Instrum Meas* 2023;72:1–14.

[15] Sun Y, Du Y, Zhang Y, Yang J, Liu J, Tian R, et al. Anti-swelling and photoresponsive mxene-based polyampholyte hydrogel sensors for underwater positioning and urban waterlogging pre-warning. *J Mater Chem A* 2024;12(33):22166–79.

[16] Hassan A, Belal A, Hassan M, Farag F, Mohamed E. Potential of thermal remote sensing techniques in monitoring waterlogged area based on surface soil moisture retrieval. *J Afr Earth Sci* 2019;155:64–74.

[17] Zhao Y, Deng Y, Pan C, Guo L. Research of water hazard detection based on color and texture features. *Sens Transducers* 2013;57(10):428–33.

[18] Chen J, Zhao N, Zhang R, Chen L, Huang K, Qiu Z. Refined crack detection via LECSFormer for autonomous road inspection vehicles. *IEEE Trans Intell Veh* 2023;8(3):2049–61.

[19] Prasad MG, Chakraborty A, Chalasani R, Chandran S. Quadcopter-based stagnant water identification. In: *Proceedings of the 2015 Fifth National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics (NCVPRIPG)*; 2015 Dec 16–19; Patna, India. Piscataway: IEEE; 2015. p. 1–4.

[20] Kim J, Baek J, Choi H, Kim E. Wet area and puddle detection for advanced driver assistance systems (ADAS) using a stereo camera. *Int J Control Autom Syst* 2016;14(1):263–71.

[21] Zhao J, Wu H, Chen L. Road surface state recognition based on SVM optimization and image segmentation processing. *J Adv Transp* 2017;2017:6458459.

[22] Yang HJ, Jang H, Jeong DS. Detection algorithm for road surface condition using wavelet packet transform and SVM. In: *Proceedings of the 19th Korea–Japan Joint Workshop on Frontiers of Computer Vision*; 2013 Jan 30–Feb 1; Incheon, Republic of Korean. Piscataway: IEEE; 2013. p. 323–6.

[23] Dias TM, Alves V, Alves H, Pinheiro L, Pontes R, Araujo G, et al. Autonomous detection of mosquito-breeding habitats using an unmanned aerial vehicle. In: *Proceedings of the 2018 Latin American Robotic Symposium, 2018 Brazilian Symposium on Robotics (SBR) and 2018 Workshop on Robotics in Education (WRE)*; 2018 Nov 6–10; João Pessoa, Brazil. Piscataway: IEEE; 2018. p. 351–6.

[24] Du YL, Yi TH, Li XJ, Rong XL, Dong LJ, Wang DW, et al. Advances in intellectualization of transportation infrastructures. *Engineering* 2023;24:239–52.

[25] Zhu B, Xia X, Wu Z. Evolutionary game theoretic demand-side management and control for a class of networked smart grid. *Automatica* 2016;70:94–100.

[26] Adugna A, Xu W, Fan J. Comparison of random forest and support vector machine classifiers for regional land cover mapping using coarse resolution FY-3C images. *Remote Sens* 2022;14(3):574.

[27] Spencer BF, Hoskere V, Narazaki Y. Advances in computer vision-based civil infrastructure inspection and monitoring. *Engineering* 2019;5(2):199–222.

[28] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: *Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017)*; 2017 Dec 4–9; Long Beach, CA, USA. New York City: Curran Associates Inc.; 2017. p. 6000–10.

[29] Yuan C, Li L, Xia X, Xiong D, Li Y, Hu J, et al. Enhancing road safety: real-time classification of low visibility foggy weather using ABNet deep-learning model. *J Transp Eng Part A Syst* 2024;150(10):04024060.

[30] Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation. In: Navab N, Hornegger J, Wells WM, Frangi AF, editors. *Medical image computing and computer-assisted intervention–MICCAI 2015*. Cham: Springer; 2015. p. 234–41.

[31] Zhang R, Yang S, Lyu D, Wang Z, Chen J, Ren Y, et al. AGSENet: a robust road ponding detection method for proactive traffic safety. *IEEE Trans Intell Transp Syst* 2024;26(1):497–516.

[33] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D. An image is worth 16x16 words: transformers for image recognition at scale. 2020. arXiv:2010.11929v2.

[34] Wan Q, Huang Z, Lu J, Yu G, Zhang L. Seaformer: squeeze-enhanced axial transformer for mobile semantic segmentation. In: Proceedings of the Eleventh International Conference on Learning Representations; 2023 May 1–5; Kigali, Rwanda. Appleton: ICLR; 2023. p. 1–19.

[35] Li Z, Wang C, Liao H, Li G, Xu C. Efficient and robust estimation of single-vehicle crash severity: a mixed logit model with heterogeneity in means and variances. Accid Anal Prev 2024;196:107446.

[36] Chen L, Yang J, Kong H. Lidar-histogram for fast road and obstacle detection. In: Proceedings of the 2017 IEEE International Conference on Robotics and Automation (ICRA); 2017 May 29–Jun 3; Singapore. Piscataway: IEEE; 2017. p. 1343–8.

[37] Colace L, Santoni F, Assanto G. A near-infrared optoelectronic approach to detection of road conditions. Opt Lasers Eng 2013;51(5):633–6.

[38] Kawai S, Takeuchi K, Shibata K, Horita Y. A method to distinguish road surface conditions for car-mounted camera images at night-time. In: Proceedings of the 2012 12th International Conference on ITS Telecommunications; 2012 Nov 5–8; Taipei, China. Piscataway: IEEE; 2012. p. 668–72.

[39] Hou Y, Li Q, Zhang C, Lu G, Ye Z, Chen Y, et al. The State-of-the-Art review on applications of intrusive sensing, image processing techniques, and machine learning methods in pavement monitoring and analysis. Engineering 2021;7(6):845–56.

[40] Yin H, Zheng F, Duan HF, Savic D, Kapelan Z. Estimating rainfall intensity using an image-based deep learning model. Engineering 2023;21:162–74.

[41] Fani SV, Kamirul, Noer A, Bissa SYC. U-Net based water region segmentation for LAPAN-A2 MSI. In: Proceedings of the 2022 IEEE International Conference on Aerospace Electronics and Remote Sensing Technology (ICARES); 2022 Nov 24–25; Yogyakarta, Indonesia. Piscataway: IEEE; 2022. p. 1–5.

[42] Han X, Nguyen C, You S, Lu J. Single image water hazard detection using FCN with reflection attention units. In: Ferrari V, Hebert M, Sminchisescu C, Weiss Y, editors. Computer vision—ECCV 2018. Cham: Springer; 2018. p. 105–21.

[43] Yang X, del Rey Castillo E, Zou Y, Wotherspoon L, Yang J, Li H. Automated concrete bridge damage detection using an efficient vision transformer-enhanced anchor-free yolo. Engineering 2025;51:311–26.

[44] Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial networks. 2014. arXiv:1406.2661.

[45] He K, Sun J, Tang X. Single image haze removal using dark channel prior. In: Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition; 2009 Jun 20–25; Miami, FL, USA. Piscataway: IEEE; 2009. p. 1956–63.

[46] Li B, Peng X, Wang Z, Xu J, Feng D. AOD-Net: all-in-one dehazing network. In: Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV); 2017 Oct 22–29; Venice, Italy. Piscataway: IEEE; 2017. p. 4780–8.

[47] Lee S, Son T, Kwak S. FIFO: learning fog-invariant features for foggy scene segmentation. In: Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2022 Jun 18–24; New Orleans, LA, USA. Piscataway: IEEE; 2022. p. 18889–99.

[48] Ma X, Wang Z, Zhan Y, Zheng Y, Wang Z, Dai D, et al. Both style and fog matter: cumulative domain adaptation for semantic foggy scene understanding. In: Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2022 Jun 18–24; New Orleans, LA, USA. Piscataway: IEEE; 2022. p. 18900–09.

[49] Bi Q, You S, Gevers T. Learning generalized segmentation for foggy-scenes by bi-directional wavelet guidance. In: Walsh T, Shah J, Kolter J, editors. Proceedings of the AAAI conference on artificial intelligence. Washington, DC, AAAI Press; 2024. p. 801–9.

[50] Chen Y, Dai X, Liu M, Chen D, Yuan L, Liu Z. Dynamic convolution: Attention over convolution kernels. In: Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2020 Jun 13–19; Seattle, WA, USA. Piscataway: IEEE; 2020. p. 11027–36.

[51] Han K, Wang Y, Guo J, Wu E. ParameterNet: parameters are all you need for large-scale visual pretraining of mobile networks. In: Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2024 Jun 16–22; Seattle, WA, USA. Piscataway: IEEE; 2024. p. 15751–61.

[52] Bi Q, You S, Gevers T. Generalized foggy-scene semantic segmentation by frequency decoupling. In: Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW); 2024 Jun 17–18; Seattle, WA, USA. Piscataway: IEEE; 2024. p. 1389–99.

[53] Sun L, Pan J, Tang J. ShuffleMixer: an efficient convnet for image super-resolution. Adv Neural Inf Process Syst 2022;35:17314–26.

[54] Zhang Q, Barri K, Babanajad SK, Alavi AH. Real-time detection of cracks on concrete bridge decks using deep learning in the frequency domain. Engineering 2021;7(12):1786–96.

[55] Gao X, Qiu T, Zhang X, Bai H, Liu K, Huang X, et al. Efficient multi-scale network with learnable discrete wavelet transform for blind motion deblurring. In: Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2024 Jun 16–22; Seattle, WA, USA. Piscataway: IEEE; 2024. p. 2733–42.

[56] Hwang S, Han D, Jung C, Jeon M. WaveDH: wavelet sub-bands guided convnet for efficient image dehazing. 2024. arXiv:2404.01604.

Journal Pre-proofs

[57] Godard C, Mac Aodha O, Firman M, Brostow G. Digging into self-supervised monocular depth estimation. In: Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV); 2019 Oct 27–Nov 2; Seoul, Republic of Korean. Piscataway: IEEE; 2019. p. 3827–37.

[58] Wei J, Wang S, Huang Q. F³net: Fusion, feedback and focus for salient object detection. Proc Conf AAAI Artif Intell 2020;34(7):12321–8.

[59] Huang Z, Wang X, Wei Y, Huang L, Shi H, Liu W, et al. Ccnet: criss-cross attention for semantic segmentation. IEEE Trans Pattern Anal Mach Intell 2023;45(6):6896–908.

[60] Qin X, Zhang Z, Huang C, Dehghan M, Zaiane OR, Jagersand M. U²-net: going deeper with nested U-structure for salient object detection. Pattern Recognit 2020;106:107404.

[61] Xie E, Wang W, Yu Z, Anandkumar A, Alvarez JM, Luo P. Segformer: simple and efficient design for semantic segmentation with transformers. Adv Neural Inf Process Syst 2021;34:12077–90.

[62] Wang H, Cao P, Yang J, Zaiane O. Narrowing the semantic gaps in U-Net with learnable skip connections: the case of medical image segmentation. Neural Netw 2024;178:106546.

[63] Yan H, Wu M, Zhang C. Multi-scale representations by varying window attention for semantic segmentation. 2024. arXiv:2404.16573v1.

[64] Wang S, Nguyen C, Liu J, Zhang K, Luo W, Zhang Y, et al. Homography guided temporal fusion for road line and marking segmentation. In: Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV); 2023 Oct 1–6; Paris, France. Piscataway: IEEE; 2023. p. 1075–85.

Declaration of Interest Statement

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐ The author is an Editorial Board Member/Editor-in-Chief/Associate Editor/Guest Editor for this journal and was not involved in the editorial review or the decision to publish this article.

☐ The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: