A Regional-Level Resource-Saving Model for Winter Road Surface Snow Detection in Extreme Weathers

Xinhao Zhou[1], Tong Wang[1], Zhaodong Liu[1], Hao Wei[1], Guangyuan Pan[1,2,*]
[1]School of Automation and Electrical Engineering, Linyi University,Shandong 276000,China
[2]Department of Civil and Environmental Engineering,University of Waterloo,Waterloo N2L 3G1,Canada
{981851821,598292172}@qq.com,liuzhaodong2017@sina.cn,weihao@lyu.edu.cn,garrypan0512@gmail.com

Abstract

Achieving timely and accurate snow detection on road surfaces in extreme weather conditions is vital for both transportation and computer vision applications. However, conventional object detection models, particularly those designed for small targets, fall short in addressing the challenge that is posed by special regional-level multiscale recognition task. To this end, an end-to-end precise and swift road surface snow detection architecture, termed the Resource-Saving Snow Detect Model (RSSD) that includes a multidimensional directional attention mechanism, is proposed. In this model, we designed three dedicated modules, namely Multi-dimensional Bidirectional Attention Module (MDBA), Split-EMA-Convolution (SEC) and Equal Split Convolution (ESC), to address the essential feature extraction and fusion tasks in snow detection. MDBA is able to promote lateral interaction and comprehensive feature fusion across scales, while SEC can not only enhance feature extraction for regional awareness but also reduces computational load, making it efficient under minimal computational power consumption. ESC preserves feature height fusion while significantly reducing computational costs, thereby enhancing the real-time detection capability of the model. In experimental evaluations conducted with data collected by in-vehicle cameras from various roads in the United States and Canada, the results demonstrate higher detection accuracy and speed compared to the latest Transformer-based real-time object detection methods and other exiting methods in the literature.Furthermore, we validated the model's performance and data sensitivity through semi-supervised learning with 50,000 unlabeled images. This research holds significant implications for winter road traffic and provides valuable insights for similar computer vision tasks.

1. Introduction

As transportation systems and computer vision technology advance, there is a growing demand for improved road safety and efficiency. Convolutional Neural Networks [9] (CNNs) and Vision Transformers [4] (ViTs) have demonstrated outstanding performance in computer vision tasks, offering new possibilities for applications in the transportation sector.

However, traditional vision models often suffer from excessive parameters, resulting in slower inference speeds and limiting their practical applications. To address this issue, anchor-based models such as the YOLO [17] series have rapidly gained attention by swiftly predicting bounding box positions and categories. The popularity of Vision Transformers has driven the development of Transformer-based object detection algorithms like DETR [1], which simplify model structures and improve processing speeds. RT-DETR [30] achieves real-time object detection by efficiently handling multiscale features. Yet, challenges persist in handling specialized regional-level multiscale recognition tasks, particularly in scenarios like road surface snow detection, where distinguishing snow from the background remains a significant challenge and exhibits notable distinctions from conventional object detection tasks. Effective snow detection can significantly reduce traffic accidents caused by snow and ice, enhance road usage efficiency, and provide accurate information on snow removal and deicing needs for road maintenance departments, optimizing resource allocation. Furthermore, with the advancement of autonomous driving technologies, precise road surface condition recognition becomes critically important for ensuring the safety of autonomous vehicles. Therefore, exploring and evaluating road detection models specifically designed for this task has become a key research direction.

In tackling these challenges, we present an end-to-end, precision-driven, and expedient architecture for road surface snow detection, named the "Resource-Saving Snow Detection Model" (RSSD). Inspired by the design of RT-

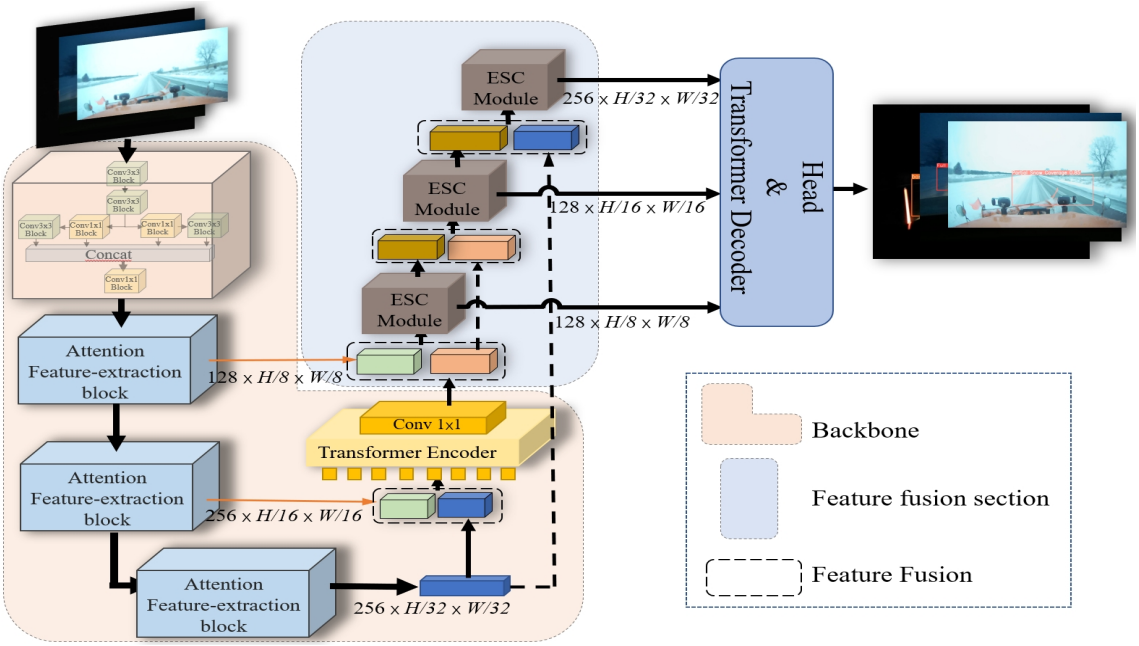

Figure 1. Network architecture diagram of RSSD

DETR, our model embraces an en-coder-decoder structure with a lightweight backbone and detection head to streamline computations and reduce the risk of overfitting. This not only ensures rapid detection but also enables the capture of nuanced road surface conditions. Within this framework, we introduce three dedicated submodules: Multi-dimensional Bidirectional Attention Module (MDBA), Split-EMA-Convolution (SEC), and Equal Split Convolution (ESC), designed to handle crucial feature extraction and fusion tasks for snow detection. The Multi-dimensional Bidirectional Attention Module employs a parallel structure, generating diverse perspectives by utilizing convolutional kernels of different sizes. Subsequently, the input features are re-evaluated from both horizontal and vertical directions, assigning higher weights to the road surface region. This parallel blending attention mechanism effectively promotes lateral interaction and comprehensive feature fusion across scales. SEC, comprising an Efficient Multi-Scale Attention Module (EMA) [14] and multiple convolutional modules, divides channels of the original input based on varying weights to retain information per channel and reduce computational burdens. In the neck segment of the model, we introduce ESC as the primary module for feature fusion. The ESC module uniformly partitions channels in the original input, applies separate convolutions, and subsequently accumulates weights. This approach preserves feature height fusion while significantly reducing computational costs, enhancing the model's real-time detection capability.

The primary contributions of this study are as follows:

(1) To address the limitations of object detection algorithms in identifying road surface conditions under extreme weather conditions, we designed a lightweight, fast, and efficient detection model and proposed an innovative multidimensional directional attention mechanism to enhance the model's capability in detecting road surface conditions.

(2) We validated our proposed innovative multidimensional directional attention mechanism using the KITTI [5] dataset and conducted comprehensive experimental validation of our road condition detection model using images captured by the Automated Vehicle Locator (AVL) system [27] in Iowa, USA, and onboard cameras in Ontario, Canada.

(3) Additionally, we further validated the model's performance and data sensitivity by employing semi-supervised learning with 50,000 unlabeled images. The experimental results demonstrate that our model outperforms traditional Transformer-based detection methods in recognizing road surface conditions, thereby confirming its superior performance and practical value.

2. Related Works

2.1. Road Surface Condition Recognition

Artificial intelligence technology has been extensively applied in the domain of intelligent transportation, particularly excelling in the recognition of road conditions. For instance, McFall and Niittula [13] utilized high-definition images specific to time and geographical location, employing artificial neural networks to detect the presence of snow

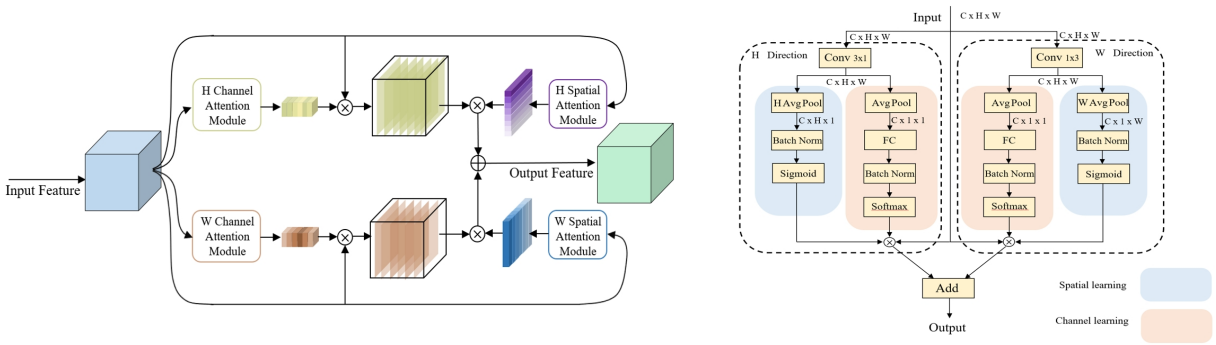

Figure 2. Illustration of our proposed Multi-Dimensional Bidirectional Attention Module Structure Diagram

and ice with an accuracy exceeding 80%. However, the model's portability was limited, rendering it unsuitable for mobile applications. Additionally, transformer back-bone networks have been applied to enhance the classification and recognition capabilities for winter road conditions [15]. Wu and colleagues [28] have further advanced this field by automating the design of CNN architectures and integrating semantic segmentation and generative adversarial network techniques. Chen et al [2]. introduced the Adaptive Hybrid Attention-based Convolutional Neural Network (AHA-CNN) framework, effectively supporting tasks such as license plate recognition and road surface monitoring.

2.2. Object Detection Methods

Although image classification methods are prevalent in research, they face challenges in real-time detection of road surface conditions due to the need for preprocessing and feature extraction before classification, which limits their responsiveness. In complex traffic scenarios, such as adverse weather conditions or high vehicle density, these methods may not capture changes in road surface conditions swiftly and accurately. Conversely, object detection algorithms in computer vision overcome these challenges by directly and efficiently capturing dynamic information about road surface states and responding quickly in real-time scenarios, making them more suitable for monitoring road conditions. The evolution of object detection from the early handcrafted feature design of the Viola-Jones [22] algorithm to deep learning approaches like the YOLO series and Faster R-CNN [6] has marked significant milestones in this field. Particularly, the YOLO algorithm, which converts detection tasks into regression problems, can predict the positions and categories of multiple objects in a single forward pass, significantly enhancing speed and accuracy. Faster R-CNN uses a Region Proposal Network for precise object localization and classification, while transformer architectures like DETR and RT-DETR improve object detection efficiency and accuracy by simplifying training processes and efficiently handling multi-scale features.

2.3. Transformer Related Methods

Transformer [21] was initially introduced in 2017, incorporating the self-attention mechanism that allows models to dynamically focus on different positions in the input sequence when processing sequential data. In 2020, it was discovered that Transformer could be applied to image-related tasks, leading to the proposal of Vision Transformer (ViT), which adopted a pure Transformer structure. ViT divides images into fixed-size blocks, treating them as sequence inputs processed by the Transformer.

Following this groundbreaking work, numerous researchers have harnessed the powerful capabilities of Transformer to extend its applications in various visual tasks. In 2021, Swin Transformer [12] was introduced, incorporating a hierarchical local attention mechanism and window-based visual block strategy to ViT. This approach maintains local information while achieving global perception, making it suitable for image tasks at various scales. In the same year, Wu et al. combined traditional convolutional neural networks (CNN) and Transformer structures, presenting the Convolutional Vision Transformer (CvT) [26] which overcame limitations of traditional CNNs in handling global information.

As Transformer research deepened, object detection models based on the Transformer architecture were continually proposed, making the application of Transformer in practical tasks feasible.

3. Proposed Method

3.1. Overall Architecture

Convolutional operations are limited in capturing local information, but by expanding the receptive field, we can more effectively process long-range relationships between pixels, which is essential for tasks such as road surface condition analysis. To enhance this capability, we have developed the RSSD model, which comprises three key modules: the Split-EMA Convolution Module (SEC), Multi-dimensional Bidirectional Attention Module (MDBA), and

the Equal Split Convolution Module (ESC), as shown in Figure 1. These modules are specifically designed to improve the accuracy and efficiency of road condition detection, with particular emphasis on optimizing feature extraction and fusion for snow detection applications.

In the RSSD model, the architecture is divided into the Backbone part and the Feature Fusion section. The Backbone part consists of a convolutional feature extraction block, three Attention Feature-extraction blocks, and a Transformer Encoder.

Within the Attention Feature-extraction Block, we have integrated our proposed MDBA and SEC. This combination allows for a deeper analysis and more meticulous extraction of input features, as illustrated in Figure 3. This approach not only enhances the model's ability to perceive important features but also significantly improves the capture of details in complex environments, thereby exhibiting exceptional performance in handling highly dynamic and variable visual scenes, such as road condition monitoring or natural environment analysis.

In the Feature Fusion section, we employ three ESC modules, which are designed to optimize and enhance the information flow between different feature layers, with a very small parameter footprint—equivalent to that of a single 3×3 convolution module. Through the coordinated work of these three ESC modules, effective deep integration of features is achieved, thus enhancing the model's ability to parse complex data structures. Each ESC module is responsible for uniformly and effectively fusing features from different network layers, not only improving the efficiency of information utilization but also enhancing the overall model's capacity to capture and respond to key information. Therefore, this carefully designed feature fusion strategy significantly elevates the model's accuracy and robustness in performing complex visual tasks, such as image segmentation and object recognition.

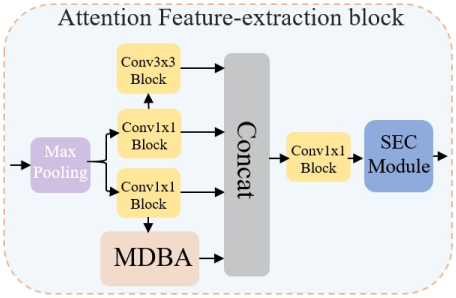

Figure 3. Attention Feature-Extraction Block Diagram

3.2. Multi-dimensional Bidirectional Attention Module

The parallel architecture of MDBA can reduce data processing and network depth, and the structure of MDBA is illustrated in Figure 2. Given the aforementioned definition of parallel processing strategy, we adopted an H and W parallel network structure in our MDBA to handle specific directional feature extraction.

Specifically, to enhance the expression capability of input features in different directions on channel and spatial dimensions, two different directional convolution kernels, namely 3×1 and 1×3, are used to capture features in the vertical and horizontal directions of the input feature map. According to the local receptive field of the convolution kernel, neurons can collect multi-scale directional information. To capture dependencies between all dimensions and reduce computational costs, in the two pathways, preliminary feature extraction in the H and W directions is performed using 3×1 and 1×3 convolution kernels, respectively. Then, two parallel pathways perform feature extraction in channel and spatial dimensions separately. Therefore, the designed MDBA is able to extract attention weight descriptors of input feature maps through two parallel pathways.

In the channel dimension, a 2D global average pooling operation and a fully connected layer are performed to enhance local cross-channel interaction, expand feature space, and perform channel recalibration through batch normalization layer to enhance features in certain channels. Finally, a SoftMax layer highlights the description capabilities of different channels for H and W direction features. The 2D global pooling operation is formulated as

$$F_{ch} = \frac{1}{H \times W} \sum_{h=1}^{H} \sum_{w=1}^{W} A_{h,w}$$

The H and W are the two dimensions of the feature, and A_{hw} is the element of the feature map at the h-th row and w-th column.
The $Softmax$ operation is formulated as

$$Softmax(Ch) = \frac{e^{c_i}}{\sum_1^C e^{c_i}}$$

The C represents the number of output channels, and c_i is the feature of the i-th output channel.

In the spatial dimension, 2D pooling operations in H and W directions are applied, and a non-linear sigmoid function is used to fit the two-dimensional binomial distribution after linear convolution. Finally, element-wise weight assignment is performed on input features. Global information in different directions is learned at the channel and spatial levels. and the H and W directional feature information extracted from the two pathways is fused to obtain the final output. The formula for 2D average pooling in the H and W directions is:

$$Out_W = \frac{1}{W} \sum_{i=1}^{W} F_{h,w} \quad \text{for } i = 1, 2, \ldots, H$$

$$Out_H = \frac{1}{H}\sum_{i=1}^{H} F_{h,w} \quad \text{for } i = 1, 2, \ldots, W$$

The $Sigmoid$ operation is formulated as

$$Sp_{(H(W))} = \frac{1}{1 + e^{-h(w)_i}}$$

The $h(w)_i$ is the i-th spatial directional feature along the H (W) direction.

3.3. Split-EMA-Convolution Module

The SEC module incorporates the Efficient Multi-Scale Attention (EMA) mechanism [14] into the channel attention and integrates multiple convolutional modules. The EMA module aims to more efficiently preserve channel information, reducing computational costs.

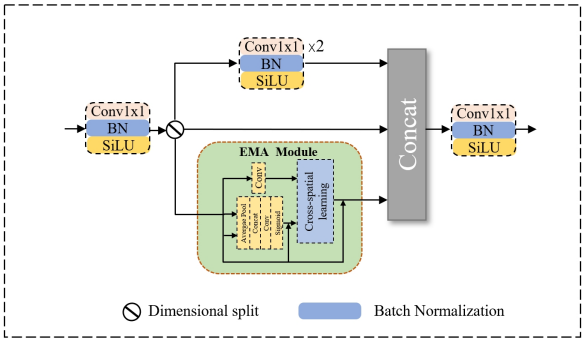

Figure 4. The Split-EMA-Convolution Module

This module reshapes partial channels into a batch dimension, splitting the channel dimension into multiple sub-features for well-distributed spatial semantic distribution within each feature group. By recalibrating channel weights within each parallel branch and further aggregating the output features of the two parallel branches through global information encoding, the module captures pixel-level pairwise relationships. The specialized design of the module achieves channel-wise splitting and attention, preserving information for each channel and achieving regional awareness through cost-effective computation across multiple sub-features. Figure 4 depicts the SEC module. The SEC module's ad-vantage lies in uniformly distributing surface features, reducing computational load through joint attention and convolution calculations on the original channels, thereby enhancing perception of road surface state changes.

3.4. Equal Split Convolution Module

The Equal Split Convolution Module (ESC) is designed as the core module of the model's neck section. The ESC module evenly divides the original input channels, subjecting the first half of the divided channels to 1×1 convolution operations and 3×3 convolution operations, while the second half undergoes channel splitting and is then subjected to two 3×3 convolution operations before being accumulated. The feature weights obtained from the two parts are then highly fused through weight accumulation. Figure 5 illustrates the structure of the ESC Module. The module's design not only maintains high-feature fusion but significantly reduces computational costs, enhancing the model's efficiency in real-time detection tasks. The ESC module aims to maintain the comprehensiveness and balance of features, providing a more effective feature fusion mechanism for the model in road surface state detection tasks.

Three ESC modules are used in the feature fusion stage of the model, respectively merging features from inputs of different sizes and outputting features with dimensions of $H \times W$, being 80×80, 40×40, and 20×20, catering to the detection of small, medium, and large dimensions of targets.

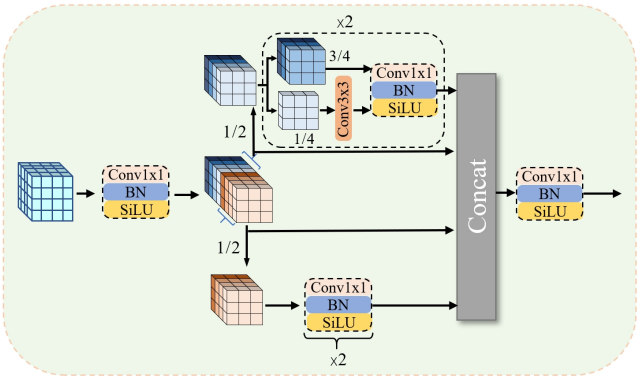

Figure 5. The Equal Split Convolution Module

4. Experimental Analysis

4.1. Experimental Design for MDBA

In this section, we present the details of our experiments and results to demonstrate the performance and efficiency of our proposed MDBA. We conducted experiments on challenging computer vision tasks. To assess the performance of MDBA, we employed the PyTorch [16], YOLOv8 standard network architecture based on Ultralytics as the baseline model. MDBA, along with four other attention mechanisms—SE [8], CA [7], ECA [24], CBMA [25]—were integrated into this architecture. We utilized the open-source KITTI dataset [5] for validation. All experiments were conducted on a PC equipped with an RTX 3060 GPU and an Intel(R) i7-11700 CPU@2.50GHz running a Windows operating system.

Table 1. Experimental Results for Road Surface Condition Detection

Models	Datasets	RT-DETR-L	RT-DETR-X	YOLOv7	YOLOv8	YOLOv8s	RT-DETR-R18	RSSD (ours)
AP_{50}^{test}	US	**88.4**	84.9	85.6	87.3	88.1	86.2	87.8
	Canada	86	87.1	90.4	92.7	**93.1**	90.4	89
	Mix	88.5	84.6	**91.2**	91.1	90.2	88	90
$Precision^{test}$	US	87.6	85.6	85.6	87.1	88.1	86.5	**88.2**
	Canada	88.3	87	84.6	87.7	86.2	90.4	**90.3**
	Mix	88.2	87.8	84.5	88.5	86.2	90.3	**91.1**
$Recall^{test}$	US	86.4	84.5	86.5	86.7	85.5	86.3	**87.3**
	Canada	86.3	87.5	86.8	**89.8**	88.3	86.4	88.2
	Mix	86.4	84.6	88	88.7	88.1	86.5	**88.8**
$Params$	——	31.9	65.4	36.5	76.7	11.1	19.8	**7**
$GFLOPs$	——	103.4	222.5	103.2	191.9	28.7	56.9	**16.8**

Table 2. Performance Exhibition of Various Attention Mechanisms on the KITTI Dataset

Models	AP_{50-95}	AP_{50}	Precision	Recall	Params
YOLOV8	77	94.9	95.3	89.5	76
+SE	76.9	94.7	94.6	89.2	77
+ECA	76.8	94.4	**95.6**	89.5	77
+CBAM	76.8	94.7	94.2	89.8	77
+MDBA(ours)	**78.2**	**95.3**	94.7	**90.9**	79

4.1.1 Dataset Description

We validated the effectiveness of MDBA using the traffic detection KITTI dataset [5], a widely utilized resource in autonomous driving and computer vision research, consisting of a large collection of images captured in real-world road scenarios. Specifically, we utilized its 2D object detection subset, comprising 7481 images with corresponding labels. Among these images, we designated 5984 images along with their corresponding labels as the training set, while the remaining 1497 images served as the validation set. This allowed us to compare the performance of MDBA against classical attention mechanisms.

4.1.2 Experimental Design and Results Analysis

In our experiments, we augmented the backbone of the baseline network with three layers of MDBA and four other attention mechanisms for comparative analysis. We evaluated commonly used metrics in object detection algorithms, including mean average precision (mAP), precision, and recall, as benchmarks for comparison, using the YOLO data format for presentation. The experimental results are presented in Table 2. Results demonstrate that SE, CA, and CBAM achieved performance comparable to the baseline, while ECA exhibited slightly lower performance compared to the baseline. In contrast, MDBA outperformed the baseline. MDBA significantly enhances the model's ability to select and extract global features, further affirming the extensive application potential of MDBA in tasks related to intelligent transportation and other domains.

4.2. Experiment on Road Surface Condition Detection

4.2.1 Dataset Description

U.S. In-Vehicle Dataset [27] comprises 4,324 images collected by the Automatic Vehicle Location (AVL) system in Iowa, strategically allocated for training (3,465 images), testing (409 images), and validation (450 images). Data is systematically gathered every 5-10 minutes by AVL systems equipped with in-vehicle cameras. **Ontario Canada In-Vehicle Dataset** contains 4,726 images captured by professional vehicle-mounted cameras in Ontario, designated for training (3,421 images), testing (493 images), and validation (812 images). These images document winter road conditions on Canadian highways, notably including a significant number of nighttime images.

4.2.2 Evaluation Metrics and Implementation Details

Detailed categorization of road surface conditions was performed using annotation tools, primarily focusing on detecting the presence of snow in the images. Road surface conditions in the image data were classified into three categories: 'Bare Pavement,' 'Full Snow Coverage,' and 'Partial Snow Coverage.' We employed datasets formatted in YOLO and assessed using the COCO-style Average Precision (AP) [11], with AP50, precision, and recall as primary indicators, while also taking into account model parameters. Additionally, a comparative analysis of model parameters and computational efficiency was conducted to evaluate the efficiency and practicality of our methods.

Table 3. Experimental Results for Semi-Supervised Learning

Models	Training Samples(Teacher Model)	AP_{50}^{test}	AP_{50-95}^{test}	$Precision^{test}$	$Recall^{test}$
Teacher Model	1222	36.9	77.1	81.9	76.7
	2032	51.9	85.4	87.4	86
	3057	52.2	84.8	87.6	84.7
	4064	53.5	85.4	88.3	90
Student Model	1222	38.4	76.2	81.9	76.9
	2032	52.4	86.3	88.9	87.3
	3057	55.6	87	88.8	87.8
	4064	56.2	88.3	90.3	87.4

We implemented all methods using the Ultralytics library in PyTorch [16]. In an environment utilizing an NVIDIA GeForce 3060 GPU, we adjusted the input image size to 640x640. The model optimization employed the AdamW optimizer, with a momentum set to 0.9 and weight decay set to 0.0001. Notably, training performance is closely related to batch size, leading us to choose a larger batch size of 8 to minimize interference and maximize GPU efficiency. All evaluated models underwent 150 epochs of training with an initial learning rate of 0.001. Except for DETR and conditional DETR, which utilized pre-trained weights, the remaining models did not employ pre-trained weights.

4.2.3 Evaluating Model Accuracy and Robustness in Diverse Conditions

In the context of onboard cameras, we evaluated four real-time object detection models, using the RT-DETR model as a baseline, to explore the superiority of our proposed model in road surface condition recognition. All models adopted YOLO data formats, and we conducted a comparative analysis of popular real-time object detection models including RTDETR-L, RTDETR-X,RTDETR-R18,YOLOv7 [23], YOLOv8 [18], and YOLOv8s, as well as our proposed model. The analysis highlighted the differences in detection accuracy among the models. To validate model performance, we utilized the Iowa, U.S., image dataset with an initial resolution of 1920x1080 and the Ontario, Canada image dataset with a resolution of 800x600, which includes a significant number of nighttime images, thereby enhancing the model's ability to handle real-world scenarios. Additionally, we created a mixed dataset comprising 6917 images from both the U.S. and Canada, employed to test the robustness of the model under varying resolutions and lighting conditions. The use of these datasets not only improved the model's ability to process multiscale information but also involved optimizing weights through the training and validation sets, with the final experimental results presented collectively in Table 1.

We have summarized the research in the field of road surface condition recognition, compiling experimental data from these studies and comparing it with our own experimental data. The results are presented in Table 4.

Table 4. Summary of experimental results in the research field

Models	IOU	Precision	Recall
Gai-ReLU [3]	—	94.9	—
DUNET1 [10]	79.3	87.1	67.3
dual-stream CNN [29]	—	89.8	89.6
Mobile Net [19]	—	87.3	86.9
ASPP Module [20]	86	—	—
RSSD (ours)	90	91.1	88.8

4.2.4 Model Ablation

We conducted a series of ablation studies to evaluate the effectiveness of each component in our proposed method, thereby validating the design of each module. In this section, all experiments were carried out using a mixed dataset to comprehensively demonstrate the contribution of each module to the task of snow-covered road detection.

To thoroughly validate the effectiveness of the proposed ME3D and ESC modules, we conducted ablation experiments within the MFDSD model using the mixed dataset mentioned in the previous section. Specifically, we replaced the MDBA, ESC, and SEC modules in the model with standard 3x3 convolution operations while maintaining the same number of channels, to assess the impact of reducing these modules. The results,as shown in Table 5,demonstrate the impact of removing of each component impacts the model's performance across various metrics.

4.3. Semi-Supervised Learning Experiments

In this study, we applied a semi-supervised learning approach to enhance model performance. The RSSD model was used as the teacher model, trained on varying sizes of labeled datasets from the RoadSurface Dataset, which included 1,222, 2,032, 3,057, and 4,064 samples. Addition-

(a) (b) (c) (d) (e) (f) (g)

Figure 6. RSSD detection results display, with the top row showing inputs and the bottom row showing outputs. Panels (a), (b), and (c) are from the U.S. dataset, while panels (d), (e), (f), and (g) are from the Canadian dataset.

Table 5. Ablation Experiment Results

Models	AP_{50}	Precision	Recall	Params
RSSD	90	91.1	88.8	16.8
-MDBA	90	89.4	87.8	17.1
-ESC	89.6	90	88.2	17
-SEC	89.7	89.2	87	17.1

ally, 1,832 labeled images were used as the test set, and 52,000 unlabeled images were used to generate pseudo-labels for evaluating model performance. Data augmentation, including flipping, rotations of 15-45 degrees, 90-degree rotations, and partial occlusion, was applied to all images to ensure pseudo-label quality.

After generating approximately 51,000 pseudo-labels, we combined them with the original labeled datasets to form an extended dataset. The RSSD model was then retrained as the student model on this extended dataset for 100 epochs.

This semi-supervised learning strategy effectively utilized a large amount of unlabeled data, significantly improving the model's generalization under limited labeled data. As shown in Table 3, the use of pseudo-labels enhanced model accuracy in the target detection task without substantially increasing labeling costs. The results, visualized in Figure 6, confirm the effectiveness of this approach in road surface condition detection.

4.3.1 Transformer-based Object Detection Algorithm

Table 6. Precision Demonstration of Different Transformer-Based Object Detection Algorithms

Models	$AP_{50}^{U.S}$	AP_{50}^{CAN}	AP_{50}^{mixed}	$Params$
DETR(R50)	86.2	91.0	92.2	43.2
Conditional DETR(R50)	90.0	93.5	94.3	43.2
RSSD (ours)	89.4	90.7	92.9	7.0

Our proposed RSSD model utilizes the Transformer framework integrated with convolutional neural networks for feature extraction and fusion. Compared to traditional Transformer-based object detection methods such as DETR and conditional DETR, RSSD demonstrates similar performance in detecting road surface conditions but with only one-sixth of the model parameters, and its training and inference speeds are significantly faster than DETR. The model has been validated on datasets from the United States, Canada, and a mixed dataset, undergoing 100 training epochs per dataset. The results (Table 6) demonstrate the innovation and effectiveness of RSSD in object detection tasks, offering a viable alternative to existing methods. Comparative analysis with two other Transformer-based algorithms further highlights the performance advantages of RSSD.

5. Conclusion

This paper presents an end-to-end, efficient architecture for road surface snow detection called the Resource-Saving Snow Detection Model (RSSD). The model incorporates a multi-dimensional bidirectional attention mechanism, enhancing feature extraction from various dimensions and spatial orientations. The framework includes three key modules: Split-EMA Convolution (SEC), Multi-Dimensional Bidirectional Attention (MDBA), and Equal Split Convolution (ESC), each addressing critical feature extraction and fusion tasks.

Evaluation of models using in-vehicle camera data from winter roads in the U.S. and Canada shows that our method performs similarly to traditional real-time object detection algorithms under high-resolution and well-lit conditions. In low-resolution and nighttime settings, while our model slightly underperforms compared to convolution-based models, it still surpasses transformer-based models. Additionally, semi-supervised learning experiments reveal significant performance gains and robust generalization, particularly with limited labeled data. Overall, our model demonstrates strong accuracy and competitiveness against medium to large-scale transformer-based object detection algorithms.

References

[1] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 1

[2] Qili Chen, Guangyuan Pan, Lin Zhao, Junfang Fan, Wenbai Chen, and Ancai Zhang. An adaptive hybrid attention based convolutional neural net for intelligent transportation object recognition. *IEEE Transactions on Intelligent Transportation Systems*, 2022. 3

[3] Lushan Cheng, Xu Zhang, and Jie Shen. Road surface condition classification using deep learning. *Journal of Visual Communication and Image Representation*, 64:102638, 2019. 7

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1

[5] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3354–3361. IEEE, 2012. 2, 5, 6

[6] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015. 3

[7] Qibin Hou, Daquan Zhou, and Jiashi Feng. Coordinate attention for efficient mobile network design. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13713–13722, 2021. 5

[8] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. 5

[9] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012. 1

[10] Caojia Liang, Junfeng Ge, Wei Zhang, Kang Gui, Faouzi Alaya Cheikh, and Lin Ye. Winter road surface status recognition using deep semantic segmentation network. In *Proceedings of the International Workshop on Atmospheric Icing of Structures (IWAIS 2019), Reykjavik, Iceland*, pages 23–28, 2019. 7

[11] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 6

[12] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 3

[13] Kevin McFall and T Niittula. Results of av winter road condition sensor prototype. In *International Road Weather Congress (SIRWEC)*, 2002. 2

[14] Daliang Ouyang, Su He, Guozhong Zhang, Mingzhu Luo, Huaiyong Guo, Jian Zhan, and Zhijie Huang. Efficient multi-scale attention module with cross-spatial learning. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. 2, 5

[15] Guangyuan Pan, Zhiyuan Bai, Liping Fu, Lin Zhao, and Qingguo Xiao. Road meteorological state recognition in extreme weather based on an improved mask-rcnn. In *International Conference on Neural Information Processing*, pages 3–15. Springer, 2023. 3

[16] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017. 5, 7

[17] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016. 1

[18] Dillon Reis, Jordan Kupec, Jacqueline Hong, and Ahmad Daoudi. Real-time flying object detection with yolov8. *arXiv preprint arXiv:2305.09972*, 2023. 7

[19] Tomoyuki Takase, Sho Takahashi, Toru Hagiwara, Tomonori Ohiro, Yuji Iwasaki, Teppei Mori, and Yasushi Hanatsuka. An estimation method of road surface condition on winter expressway via mobile nets using in-vehicle camera images. In *2021 IEEE 10th Global Conference on Consumer Electronics (GCCE)*, pages 109–110. IEEE, 2021. 7

[20] Sirawich Vachmanus, Ankit A Ravankar, Takanori Emaru, and Yukinori Kobayashi. Semantic segmentation for road surface detection in snowy environment. In *2020 59th Annual Conference of the Society of Instrument and Control Engineers of Japan (SICE)*, pages 1381–1386. IEEE, 2020. 7

[21] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 3

[22] Paul Viola and Michael Jones. Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001*, volume 1, pages I–I. Ieee, 2001. 3

[23] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7464–7475, 2023. 7

[24] Qilong Wang, Banggu Wu, Pengfei Zhu, Peihua Li, Wangmeng Zuo, and Qinghua Hu. Eca-net: Efficient channel attention for deep convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11534–11542, 2020. 5

[25] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In

Proceedings of the European conference on computer vision (ECCV), pages 3–19, 2018. 5

[26] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. Cvt: Introducing convolutions to vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 22–31, 2021. 3

[27] Mingjian Wu, Tae J Kwon, et al. An automatic architecture designing approach of convolutional neural networks for road surface conditions image recognition: Tradeoff between accuracy and efficiency. *Journal of Sensors*, 2022, 2022. 2, 6

[28] Huynh N. Wu M, Kwon T J. Winter road surface condition recognition using semantic segmentation and the generative ad-versarial network: A case study of iowa, usa. *Transportation Research Record*, 2022. 3

[29] Ce Zhang, Ehsan Nateghinia, Luis F Miranda-Moreno, and Lijun Sun. Winter road surface condition classification using convolutional neural network (cnn): visible light and thermal image fusion. *Canadian Journal of Civil Engineering*, 49(4):569–578, 2022. 7

[30] Yian Zhao, Wenyu Lv, Shangliang Xu, Jinman Wei, Guanzhong Wang, Qingqing Dang, Yi Liu, and Jie Chen. Detrs beat yolos on real-time object detection. *arXiv preprint arXiv:2304.08069*, 2023. 1