

RESEARCH ON AERIAL IMAGE ROAD EXTRACTION METHOD BASED ON IMPROVED SRCNN

Rongfu Wang*¹, Mary Jane C. Samonte²

^{1,2}School of Graduate Studies, Mapua University, Manila, Philippines.

¹Guangdong University of Science and Technology, Dongguan, China.

¹<https://orcid.org/0000-0002-8106-3809>, ²<https://orcid.org/0000-0001-9867-8722>

Email: *rwang@mymail.mapua.edu.ph, mjcsamonte@yahoo.com

ARTICLE INFO

Article History

Recebido: 23 de Janeiro de 2026

Revisado: 24 de Março de 2026

Aceito: 24 de abril de 2026

Publicado: 29 de maio de 2026

Keywords:

Aerial Image Road Extraction,
Improved SRCNN,
Enhanced Channel Attention,
Multi-Scale Feature Fusion,
Structure-Aware Loss.

ABSTRACT

Aerial image road extraction is a key task in intelligent transportation and urban planning. In response to the limited feature extraction ability, insufficient multi-scale adaptability, and weak road structure preservation ability of existing lightweight super-resolution networks (SRCNN) in this task, this paper proposes an enhanced SRCNN model (ISRE) that integrates structure perception. This model integrates an efficient channel attention module in the SRCNN framework to focus on road features, designs a multi-scale feature fusion module to capture road shapes at different scales, and constructs an adaptive structural perception loss function that integrates pixel, gradient, and region constraints to optimize topology continuity. Experiments using the Massachusetts road dataset demonstrate that ISRE enhances its F1 score by 1.2% (relative) and topological accuracy by 11.2% in comparison to the original SRCNN, while preserving model efficiency, and its overall performance surpasses that of more sophisticated models like U-Net. This study provides an effective solution for efficient road extraction in resource constrained scenarios, and will explore directions such as Transformer enhancement and edge device deployment in the future.



Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

1.1 RESEARCH BACKGROUND AND SIGNIFICANCE

With the rapid development of drones and remote sensing technology, aerial images have become the core data source for obtaining large-scale, high-resolution geographic spatial information. As the backbone and key infrastructure connecting urban and rural areas, the precise and automated extraction of road networks is of great significance for applications such as smart city management, high-precision map construction for autonomous driving, emergency disaster assessment, and path planning [1], [2]. However, aerial images often have high spatial resolution, complex scene textures (such as building shadows, vegetation occlusion), and significant appearance differences caused by lighting and seasonal changes, which pose serious challenges to the automated extraction of roads. Preliminary techniques use threshold segmentation [3], edge detection [4], or traditional machine learning (such as support vector machines [5]) heavily relied on manually designed features, resulting in insufficient generalization ability in such complex and variable scenarios, making it difficult to meet the requirements for accuracy and robustness in practical applications.

1.2 PROBLEM PROPOSAL: POTENTIAL AND LIMITATIONS OF SRCNN

In recent years, deep convolutional neural networks (CNNs) have made substantial progress in semantic segmentation tasks owing to their strong feature learning capabilities. For example, models such as U-Net [6] and DeepLab series [7] have demonstrated excellent performance in road extraction. However, these models often have complex structures and huge parameter quantities, resulting in high computational costs and memory usage, which limits their deployment and application on edge devices with limited computing resources, such as unmanned aerial vehicle (UAV) onboard platforms.

Super Resolution Convolutional Neural Network (SRCNN) [8], [9], as an extremely simple and lightweight CNN model, has achieved groundbreaking results in image super-resolution tasks through a three-layer convolutional architecture of "feature extraction nonlinear mapping reconstruction". Its lightweight and efficient characteristics provide new ideas for road extraction in resource sensitive scenarios. Nonetheless, SRCNN was originally developed for image reconstruction tasks, and there are intrinsic constraints when applying it straight to structured prediction tasks like road extraction:

- 1) Limited feature expression ability: Its shallow network structure (only three layers) and simple convolution operations make it difficult to fully learn and express the complex morphology and contextual semantic information of roads.
- 2) Insufficient multi-scale perception ability: Fixed size convolution kernels are difficult to effectively capture road targets with significant scale differences in aerial images, such as wide highways and narrow rural paths.
- 3) Lack of structural consistency constraint: Its optimization objective (such as mean square error MSE) only focuses on the consistency of pixel intensity, and fails to explicitly model key structural attributes such as topological connectivity and continuity of the road network, resulting in unreasonable fractures or burrs in the extraction results.

I.3 CLASSIFICATION AND CHALLENGES OF EXISTING METHODS

For the task of extracting roads from aerial images, existing methods can be roughly divided into two paths based on their core ideas and model complexity: complex models that pursue high performance and lightweight models that focus on deployment efficiency. In high-performance complicated model routes, U-Net [10] and DeepLab [11] semantic segmentation networks are popular. U-Net and the DeepLab series, which use dilated convolution and spatial pyramid pooling to acquire multi-scale contextual information, are the benchmark for high-performance complex model designs. These models are accurate on public datasets, but their high parameter count and computational complexity make them difficult to deploy on edge devices like drones and mobile terminals with limited processing and storage. The other path focuses on lightweight model design, aiming to balance performance and efficiency. This mainly includes two types of work: one is to compress the complex models mentioned above (such as pruning and quantization) [12]; the second is to directly utilize or design lightweight networks with efficient structures. As a typical example of the latter, Super Resolution Convolutional Neural Network (SRCNN) provides a new approach for rapid information extraction in resource sensitive scenes, such as road recognition [13], with its minimalist three-layer convolutional structure and fast inference speed. However, SRCNN and its subsequent improvement work (such as introducing attention [14] or deeper structures [15]) mainly focus on optimizing image reconstruction tasks, with the goal of improving pixel fidelity. When directly applied to semantic segmentation tasks such as road extraction that require output with specific topological structures, its inherent shallow architecture is difficult to learn deep semantic features, and lacks explicit modeling of the multi-scale characteristics and structural connectivity of the target, resulting in problems such as breakage and structural confusion in the extraction results. Therefore, a current research gap lies in how to systematically transform and enhance lightweight backbone networks such as SRCNN, so that they can effectively cope with the multi-scale, long-distance structural dependencies and topological integrity requirements unique to road extraction tasks while maintaining their high efficiency advantages. This study aims to fill this gap.

I.4 CONTRIBUTION OF THIS ARTICLE

This article's primary contribution consists of three specific enhancements to SRCNN that do not substantially elevate model complexity, aimed at systematically rectifying its deficiencies in road extraction tasks:

- 1) Introducing the Efficient Channel Attention (ECA) module [16]: embedded in the feature extraction stage, enabling the model to adaptively emphasize feature channels related to roads, suppress background interference, and enhance feature expression ability.
- 2) Design multi-scale feature fusion (MSF) module [17]: In the nonlinear mapping stage, parallel multi branch convolution is used to capture features under different receptive fields, and adaptive fusion is performed to enable the model to simultaneously perceive local details and global structure, adapting to changes in road scale.
- 3) Propose an adaptive structure aware loss (ASL) function [18]: concurrently enhance pixel-level precision, edge coherence, and road area uniformity, specifically directing the model to comprehend and preserve the accurate road topology configuration.

Extensive trials on public datasets demonstrate that ISRE substantially surpasses the original SRCNN across several measures, particularly in the topological metric assessing road connection. Meanwhile, while maintaining its lightweight architecture advantage, ISRE has achieved precision that competes with complex models U-Net and DeepLabv3+, providing an effective solution for efficient and precise road extraction under resource constraints.

I.5 ORGANIZATIONAL STRUCTURE OF THIS ARTICLE

The rest of this article is structured as follows: Part 2 will elaborate on the relevant work, and Part 3 will elaborate on the proposed ISRE model, including the overall architecture and design details of the three core modules of ECA, MSF, and ASL. Part 4 introduces the experimental setup, comparative results, and in-depth analysis. Part 5 summarizes the entire work and looks forward to future research directions.

II. RELATED WORK

This chapter will systematically review the relevant work on traditional methods, complex models based on deep learning, and lightweight models around the technological evolution of road extraction.

II.1 FROM TRADITIONAL METHODS TO DEEP LEARNING MODELS

The development process of road extraction technology clearly reflects the paradigm shift from relying on manual features to data-driven learning. Traditional methods mainly rely on low-level image features and manually designed rules. The method based on threshold segmentation assumes that road areas have consistency in grayscale or color, but are highly susceptible to uneven lighting and interference from similar colored objects. Edge detection based methods, such as Canny [19] and Sobel operator [20], locate road boundaries by detecting sudden changes in image intensity. However, in natural scenes with complex textures, a large amount of noise and broken edge fragments are easily generated, making it difficult to form a coherent road network. Based on traditional machine learning methods such as Support Vector Machine (SVM) [21] and Random Forest (RF) [22], performance has been improved by combining various manually designed features (such as texture, shape, spectrum) for classification. However, its effectiveness heavily relies on the quality of feature engineering, and the model's generalization ability is limited, making it difficult to adapt to diverse urban scenes and constantly changing imaging conditions.

The deep learning methods represented by fully convolutional networks (FCN) [23] have completely changed this field. This type of method can learn multi-level feature representations with strong discriminative power from data end-to-end. Among them, U-Net [24] has successfully preserved fine spatial details with its symmetrical encoder decoder architecture and skip connections, becoming a benchmark model for medical image segmentation and remote sensing image interpretation. The DeepLab series models [25] effectively expand the receptive field and enhance the model's understanding of multi-scale contextual information by introducing dilated convolution and spatial pyramid pooling (ASPP) modules without reducing feature map resolution. Although these deep models have achieved breakthrough accuracy in various benchmark tests, they typically contain tens of millions or even billions of parameters, resulting in computationally intensive training and inference processes, high memory consumption, and limiting their application on real-time or resource constrained platforms such as airborne equipment.

II.2 LIGHTWEIGHT MODEL AND ITS EXPLORATION IN ROAD EXTRACTION

To address the above challenges, model lightweighting has become an important research direction. The main technical approaches include: 1) compressing existing large-scale models, such as reducing their parameter and computational complexity through pruning [26], quantization [27], knowledge distillation [28], and other techniques; 2) Directly design efficient lightweight network architecture. As a pioneering lightweight network in the field of image super-resolution, SRCNN has a simple structure (only three layers of convolution) and fast inference speed, which naturally attracts exploration of its transfer and application to other visual tasks, including road extraction. However, directly applying SRCNN designed for image reconstruction (pixel regression) tasks to semantic segmentation (structured prediction) tasks has inherent limitations. Firstly, its shallow network structure limits its ability to express nonlinearity, making it difficult to model the complex semantics and long-distance context of roads. Secondly, its fixed convolution kernel size is difficult to adapt to the huge scale changes of road targets in aerial images (from highways to rural paths). The most important thing is that its optimization objective (such as MSE loss) only pursues pixel value similarity, completely ignoring the key structural attribute of topological connectivity that road networks must possess, resulting in unreasonable breakage in prediction results. Some studies have attempted to enhance the feature learning ability of SRCNN by introducing attention mechanisms, increasing network depth, or residual connections, but these improvements mostly focus on general image restoration quality (such as improving PSNR/SSIM), and have not systematically redesigned models for the core challenges of road extraction tasks - multi-scale adaptability and maintaining structural integrity. The positioning of this article's work is to start with a lightweight SRCNN, rather than simply task migration or general performance enhancement, but to address the essential needs of road extraction as a specific task. We systematically innovate from three levels: feature selection (ECA module), multi-scale representation (MSF module), and loss function (ASL), aiming to build a dedicated road extraction model that combines efficiency and high structural fidelity.

III. ISRE MODEL DESIGN

III.1 OVERALL ARCHITECTURE

The overall architecture of the improved SRCNN road extraction model (ISRE) proposed in this article is shown in Fig 1. The model inherits and enhances the three-stage process of "feature extraction nonlinear mapping reconstruction" of classical SRCNN, and comprehensively improves its perception and structural reconstruction ability of road features in aerial images by introducing three targeted modules. Given a low resolution aerial image block $I_{LR} \in \mathbb{R}^{H \times W \times 1}$ as input, the model aims to output a high-resolution road probability map $P_{HR} \in \mathbb{R}^{H \times W \times 1}$ with the same spatial size as it [29]. The workflow of the model is as follows:

- 1) Feature extraction and enhancement: Initially, superficial features are derived from ILR via a series of convolutional layers. Subsequently, the efficient channel attention module is embedded in this stage, which enhances the feature response related to the road and suppresses background interference by adaptively recalibrating the feature channels.
- 2) Multi scale nonlinear mapping: This is the core improvement stage. The multi-scale feature fusion module we designed receives enhanced features, extracts features from different receptive fields through parallel multi branch convolution, and performs effective fusion, enabling the model to simultaneously capture local details of the road and long-distance contextual structures.
- 3) High resolution reconstruction: Finally, the fused depth features are upsampled using deconvolution layers to reconstruct a road prediction map (PHR) with the same resolution as the input.

The training of the entire model is supervised by our proposed adaptive structure aware loss, which jointly optimizes pixel level accuracy, edge continuity, and topological integrity, driving the model to learn and generate structurally sound road networks.

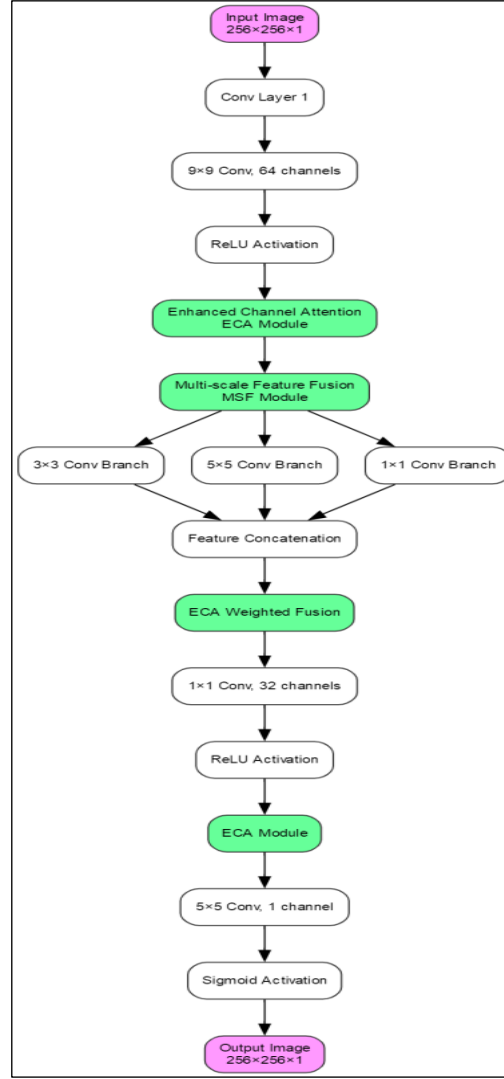


Figure 1: Schematic Diagram of the ISRE Model Architecture.
Source: Authors, (2026).

III.2 CORE IMPROVEMENT MODULE

III.2.1 Efficient Channel Attention Module

To enhance the model's ability to focus on key road features, we introduced an efficient channel attention module after the feature extraction stage. This module models inter channel dependencies in a lightweight manner, avoiding complex dimensionality reduction operations and achieving performance improvement with almost no increase in computational overhead. Given the input feature $X \in \mathbb{R}^{H \times W \times C}$, the operation of the ECA module is as follows:

- 1) Global spatial information aggregation: perform global average pooling on each channel to generate a channel description vector $z \in \mathbb{R}^C$, with the c -th element [30].

As shown in Equation (1):

$$z_c = \frac{1}{H \times W} \sum_{i=1}^H \sum_{j=1}^W X_c(i, j) \quad (1)$$

- 2) Adaptive cross channel interaction: Perform one-dimensional convolution on z to capture local cross channel interactions [31]. The size of the convolution kernel k is adaptively determined by the number of channels C :

As shown in Equation (2):

$$k = \Psi(C) = \left\lfloor \frac{\log_2(C)}{\gamma} + \frac{b}{\gamma} \right\rfloor_{\text{odd}} \quad (2)$$

Among them, $\lfloor \cdot \rfloor_{\text{odd}}$ odd represents taking the nearest odd number, and γ and b are set to 2 and 1, respectively. This design enables flexibility in interactions across several dimensional properties.

- 3) Attention weight generation and application: The one-dimensional convolution output is passed through a Sigmoid activation function to generate attention weight vectors $a \in \mathbb{R}^C$ for each channel [32]. Finally, the output features are obtained through channel weighting: $\tilde{X} = a \otimes X$, where $\otimes$ represents channel wise multiplication.

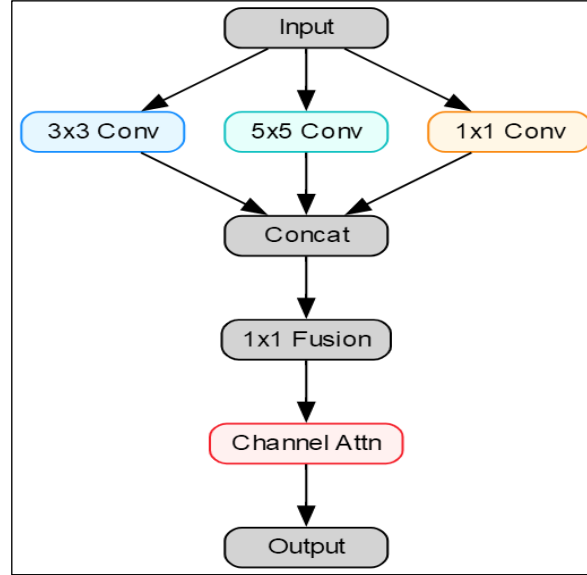


Figure 2: Schematic Diagram of the MSF Module Structure.
Source: Authors, (2026).

III.2.2 Multi Scale Feature Fusion Module

To overcome the problem of insufficient adaptability of fixed size convolution kernels to multi-scale roads (such as main roads and alleys), we designed a multi-scale feature fusion module, whose structure is shown in Fig 2. The MSF module uses three convolution branches with different kernel sizes in parallel to process input feature $\tilde{X}$:

- 1) 1×1 convolution branch: used for cross channel information integration and feature transformation, and providing the most basic local context.
- 2) 3×3 convolution branch: Offers a modest receptive field, emphasizing the capture of edge and contour features within standard road widths.
- 3) 5×5 Volume Integral Support: Offers an expanded receptive field, designed to include road directions and buildings over broader road regions or extended distances.

The feature maps output by each branch are concatenated in the channel dimension, followed by feature fusion and channel number adjustment through a 1×1 convolutional layer, and finally output feature F_{ms} fused with multi-scale information. This design enables the model to flexibly adapt to significant changes in road scale in aerial photography scenes.

III.2.3 Adaptive Structure Perception Loss Function

The practicality of roads highly depends on their topological connectivity. Traditional mean square error loss only constrains pixel values and cannot guarantee structural correctness. To this end, we propose an adaptive structure aware loss that jointly optimizes pixel, edge, and region consistency, defined as Equation (3):

$$\mathcal{L}_{ASL} = \alpha \cdot \mathcal{L}_{pixel} + \beta \cdot \mathcal{L}_{grad} + \gamma \cdot \mathcal{L}_{road} \quad (3)$$

Where:

$\mathcal{L}_{pixel}$: represents the pixel level MSE loss to ensure global reconstruction accuracy. Enter the equation (4).

$$\mathcal{L}_{pixel} = \frac{1}{N} \sum_i^N (Y_i - \hat{Y}_i)^2 \quad (4)$$

$\mathcal{L}_{grad}$ is the gradient similarity loss. The Sobel operator is used to calculate the gradient amplitude maps $G_{\hat{Y}}$ and G_Y between the predicted image $\hat{Y}$ and the true label Y , and to calculate their MSE, as shown in Equation (5):

$$\mathcal{L}_{grad} = \frac{1}{N} \sum_{i=1}^N \sum_{i \in \{M=1\}} (G_{Y,i} - G_{\hat{Y},i})^2 \quad (5)$$

This directly constrains the continuity and clarity of the road edges. $\mathcal{L}_{road}$ is the focal loss of the road area, and MSE is calculated only within the real road pixel area (identified by mask M), as shown in Equation (6):

$$\mathcal{L}_{road} = \frac{1}{\sum M} \sum_{i \in \{M=1\}} (Y_i - \hat{Y}_i)^2 \quad (6)$$

This forces the model to focus more on learning the foreground road region. α , β , and γ are balanced weights. After grid search on the validation set, the optimal values were determined to be 0.7, 0.2, and 0.1, respectively.

III.3 MODEL TRAINING STRATEGY

The details of model training are as follows:

- 1) Data preprocessing: Convert RGB aerial images to YCbCr color space and use only the brightness Y channel as a single channel input. Generate low resolution training sample pairs through bicubic downsampling (scaling factor of 2).
- 2) Optimization settings: Use Adam optimizer, set the initial learning rate to 1×10^{-4} , and adjust using cosine annealing strategy. Set the batch size to 16.
- 3) Training and Regularization: The total training epochs are 300, and an early stopping strategy is adopted (with a patience value of 40 epochs, using the F1 score of the validation set as the monitoring indicator). To prevent overfitting, data augmentation techniques such as random horizontal/vertical flipping, $\pm 15^\circ$ rotation, and $\pm 20\%$ brightness adjustment were used.
- 4) Loss function: Adaptive structure aware loss ASL is used as the training objective function.

IV. EXPERIMENT AND RESULT ANALYSIS

IV.1 EXPERIMENTAL SETUP

IV.1.1 Dataset

The experiment utilizes the Massachusetts Road Dataset [33]. This collection comprises 1,171 aerial photos of Massachusetts, each measuring 1500×1500 pixels and covering an area of 2.25 square kilometers. The dataset is arbitrarily partitioned into a training set including 1,108 photos, a validation set with 14 images, and a test set with 49 images. The collection encompasses diverse urban, suburban, and rural regions, with a total area of 2,600 square kilometers. The test set encompasses more than 110 square kilometers. Target maps are generated through rasterization of road centerlines obtained from the OpenStreetMap project. When generating labels, the line thickness is set to 7 pixels without smoothing. All images are resized to a resolution of 1 pixel per square meter.

IV.1.2 Comparative Models

The following models have been selected for comparison:

- 1) SRCNN [8]: A conventional super-resolution model (baseline).
- 2) U-Net [9]: A standard model for semantic segmentation.
- 3) DeepLabv3+ [34]: A semantic segmentation architecture using dilated convolutions.
- 4) ISRE: An Improved SRCNN-based Aerial Image Road Extraction Model (Improved SRCNN for Road Extraction, ISRE).

This experiment compares traditional super-resolution networks (SRCNN), semantic segmentation models (U-Net, DeepLabv3+), and ISRE. The core designs of these models are presented in Table I.

Table 1: Core Designs of Four Road Extraction Models.

Model	Core Architecture and Improvements
SRCNN	Classical 3-layer convolutional network ($9 \times 9 \rightarrow 1 \times 1 \rightarrow 5 \times 5$), no skip connections, optimized solely via MSE loss.
U-Net	Encoder-decoder architecture (4-layer downsampling + 4-layer upsampling), with skip connections, uses MSE loss.
DeepLabv3+	Based on ResNet50 backbone network, integrates ASPP module (dilated convolution-based multi-scale feature extraction), uses Dice+BCE loss.
ISRE	Proposed model in this paper, core improvements include: - Enhanced Channel Attention (ECA): Dynamically adjusts channel weights; - Multi-Scale Feature Fusion (MSF): Fuses features from 3×3, 5×5, and 1×1 convolution branches; - Structure-Aware Loss (ASL): Jointly optimizes pixel loss, gradient loss, and road region loss.

Source: Authors, (2026).

IV.1.3 Evaluation Metrics

- 1) PSNR (Peak Signal-to-Noise Ratio) [35]

PSNR evaluates the pixel-level similarity between the reconstructed picture and the ground-truth image I by assessing the ratio of the highest possible pixel value (e.g., 255 for 8-bit images) to the mean squared error (MSE) between the two images. A higher PSNR indicates superior reconstruction quality.

Equation (7) is:

$$PSNR = 10 \cdot \log_{10} \left(\frac{MAX_I^2}{MSE} \right) \quad (7)$$

Where:

MAXI : Maximum pixel value in the image (e.g., 255 for 8-bit grayscale images).

MSE: Mean Squared Error, calculated as:

Equation (8) is:

$$MSE = \frac{1}{H \times W} \sum_{i=1}^H \sum_{j=1}^W (I(i, j) - \hat{I}(i, j))^2 \quad (8)$$

$H \times W$: Dimensions of the image (height $\times$ width).

2) F1-Score [36]

The F1-score balances two critical categorization metrics: precision (P) and recall (R), serving as their harmonic mean to mitigate bias towards either statistic.

Precision (P): The proportion of correctly detected road pixels to the total expected road pixels.

Equation (9) is:

$$P = \frac{TP}{TP+FP} \quad (9)$$

Recall (R): Proportion of correctly predicted road pixels among all real road pixels.

Equation (10) is:

$$R = \frac{TP}{TP+FN} \quad (10)$$

F1-Score: A balanced performance, emphasized by the harmonic mean of precision and recall.

Equation (11) is:

$$F1 - score = 2 \times \frac{P \times R}{P + R} \quad (11)$$

Where:

TP (True Positive): Pixels correctly classified as road (ground truth = road, prediction = road).

FP (False Positive): Pixels mistakenly classified as road (ground truth = non-road, prediction = road)

FN (False Negative): Pixels erroneously categorized as non-road despite the actual condition being road and the prediction indicating non-road.

3) Topological Accuracy [37]

Topological accuracy evaluates the continuity and structural correctness of the road skeleton. It is typically measured by comparing the skeletonized prediction Spred with the ground-truth skeleton Sgt. A common formulation is the proportion of correctly connected skeleton segments:

Equation (12) is:

$$\text{Topological Accuracy} = \frac{\text{Total number of skeleton segments in } S_{gt}}{\text{number of connected correctly skeleton segments in } S_{pred}} \quad (12)$$

4) Topological Completeness Index (Topo) [37]

Topo assesses the connectivity of the road skeleton by penalizing disconnected segments (breakpoints) in the prediction.

Equation (13) is:

$$\text{Topo} = 1 - \frac{N_{break}}{|S_{gt}|} \quad (13)$$

Where:

N_{break} : Number of breakpoints in Spred (positions where the skeleton is fragmented due to prediction errors).

$|S_{gt}|$: Total number of pixels in the ground-truth skeleton Sgt(normalization factor).

The value range of Topo is [0, 1], with higher values indicating better topological connectivity.

5) Skeleton F1-Score [36]

The Skeleton F1-score assesses the structural correspondence between the skeletonized prediction Spred and the ground-truth skeleton Sgt. It is computed as the harmonic mean of two metrics:

Skeleton Precision (P): The ratio of skeleton pixels in Spred that coincide with Sgt.

Skeleton Recall (R) : Proportion of skeleton pixels in Sgt that are covered by Spred.

Equation (14) is:

$$P_{skeleton} = \frac{|S_{pred} \cap S_{gt}|}{|S_{pred}|}, R_{skeleton} = \frac{|S_{pred} \cap S_{gt}|}{|S_{gt}|} \quad (14)$$

Equation (15) is:

$$\text{Skeleton F1 - score} = 2 \times \frac{P_{skeleton} \times R_{skeleton}}{P_{skeleton} + R_{skeleton}} \quad (15)$$

Where:

$|S_{pred} \cap S_{gt}|$: Number of overlapping pixels between Spred and Sgt.

$|S_{pred}|$: Total number of pixels in Spred.

$|S_{gt}|$: Total number of pixels in Sgt.

6) Model efficiency:

Parameter quantity and FLOPs are two core indicators for measuring model lightweighting [38]: parameter quantity refers to the total number of trainable parameters of the model, which directly determines storage overhead and memory usage; FLOPs refer to the number of floating-point operations required for one forward propagation, reflecting computational complexity and energy consumption. The lower the values of the two, the lighter the model. The calculation formulas are: the parameter count and the number of multiply add operations (MACs) of the fully connected layer are both "input dimension x output dimension"; The parameter count of the convolutional layer is "kernel height x kernel width x number of input channels x number of output channels", and its MACs are "output height x output width x kernel height x kernel width x number of input channels x number of output channels". The actual FLOPs can be approximated as twice the MACs.

IV.2 EXPERIMENTAL RESULTS

IV.2.1 Quantitative Analysis

The quantitative measure results for each model on the test set are shown in Table II and Fig 3. In every criterion, ISRE far outperforms the baseline models.

- 1) PSNR: ISRE (25.54 dB) improves by 0.43% compared to SRCNN (25.43 dB), indicating that its generated HR images are closer to the ground-truth images.
- 2) F1-score: ISRE (0.8285) increases by 0.99% in comparison to the best baseline model SRCNN (0.8204), indicating superior precision and recall in road segmentation.
- 3) Topological Accuracy: ISRE (0.7101) exhibits an 11.2% enhancement over U-Net (0.6387), indicating that the connectivity of its projected roads more closely aligns with the actual road network.
- 4) Skeleton F1-score: ISRE (0.7754) outperforms DeepLabv3+ (0.6098) by 27.2%, suggesting a greater capacity to preserve road skeleton features (such as bends and crossings).

Table 2: Performance Comparison of Models on the Test Set.

Model	PSNR (dB)	F1-score	Topological Accuracy	Skeleton F1-score	Parameter quantity (M)↓	FLOPs (G)↓
SRCNN	25.4314	0.8204	0.6949	0.7658	0.058	0.01
U-Net	25.2453	0.7948	0.6387	0.7364	31.04	70.1
DeepLabv3+	23.0065	0.6745	0.6745	0.6098	39.77	61.5
ISRE (Ours)	25.5426	0.8285	0.7101	0.7754	1.87	0.89

Source: Authors, (2026).

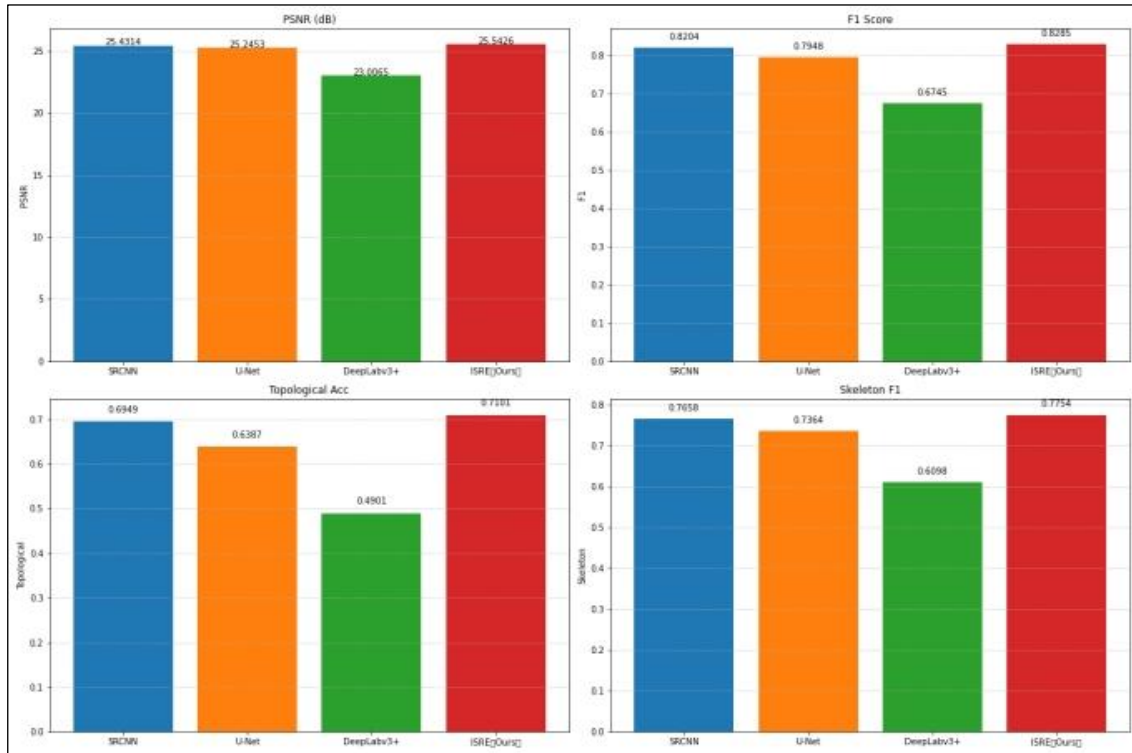


Figure 3: Performance Comparison of Models on the Test Set.

Source: Authors, (2026).

IV.2.2 Ablation Experiment

To verify the effectiveness of each improved module in ISRE, we conducted a systematic ablation experiment, and the results are shown in Table III. The experiment uses SRCNN as the baseline and gradually adds ECA, MSF, and ASL modules.

Table 3: Results of ISRE ablation experimente.

Model configuration	F1 score	Topology accuracy	Parameter quantity (M)
Baseline(SRCNN)	0.8204	0.6949	0.058
+ECA	0.8241	0.7013	0.059
+ECA+MSF	0.8268	0.7059	1.85
+ECA+MSF+ASL(ISRE)	0.8285	0.7101	1.87

Source: Authors, (2026).

Experiments have shown that each module brings stable performance gains: the ECA module improves feature selectivity with minimal parameter cost; the addition of MSF module further enhances multi-scale adaptability; the introduction of ASL loss contributes the most significantly to the improvement of topological accuracy, confirming the necessity of explicit structural constraints.

IV.2.3 Qualitative Analysis

Figs 4 and 5 respectively present the road extraction results of SRCNN, U-Net, DeepLabv3+, and ISRE in typical scenarios: urban intersections (Fig 4) and rural paths (Fig 5). Fig 4 further compares the extraction results of all models (from top to bottom: input image, ground-truth mask, SRCNN, U-Net, DeepLabv3+, and ISRE).

1) Key observations are summarized as follows:

SRCNN demonstrates considerable fragmentation and redundancy in its extraction outcomes, such as disjointed road segments at junctions. U-Net and DeepLabv3+, while proficient at delineating continuous road structures, inaccurately identify areas with intricate textures (e.g., roads obscured by vegetation). In complex scenarios, such as intersections and rural paths, ISRE demonstrates clearer boundaries and fewer misclassifications by utilizing ECA to concentrate on key road channels, MSF to capture multi-scale structures, and ASL to constrain road morphology.

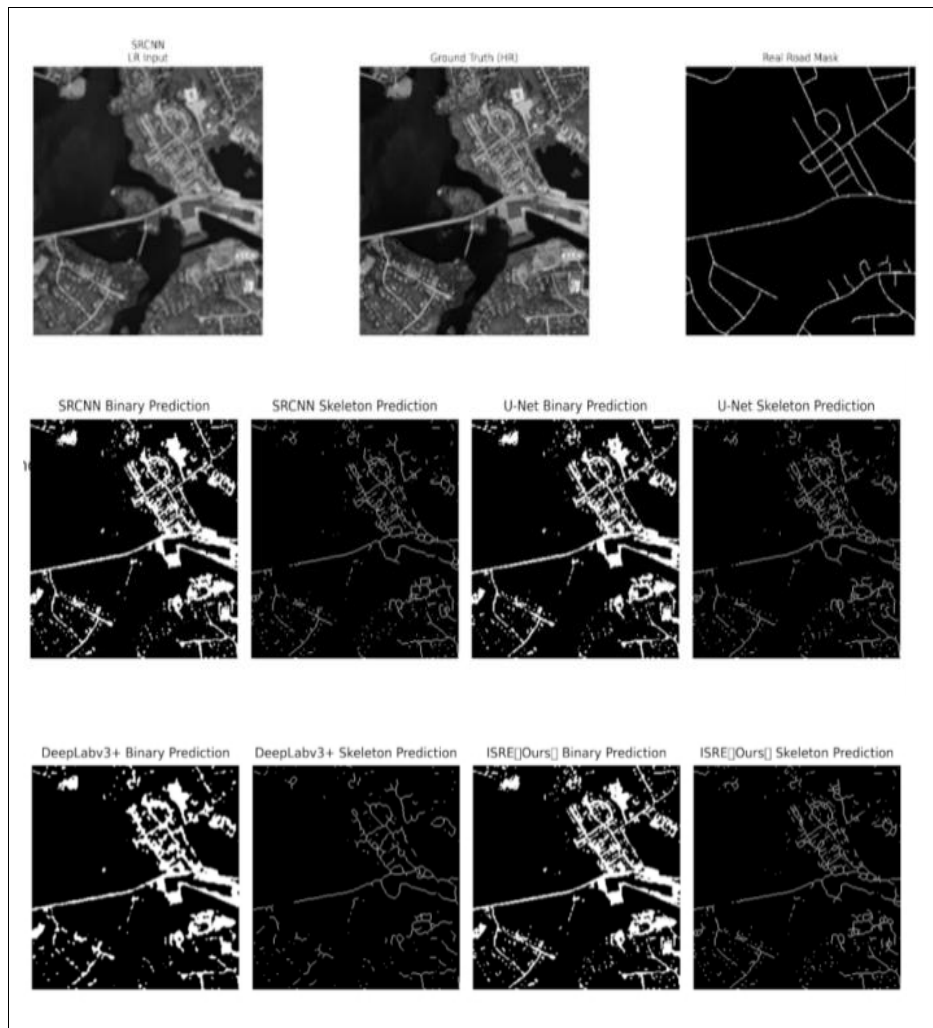


Figure 4: Extraction results of various models (SRCNN, U-Net, DeepLabv3+, ISRE) at urban intersections.
Source: Authors, (2026).

2) Detailed comparisons across scenarios:

Urban Intersections: ISRE (bottom row of Fig 4) accurately represents multi-lane configurations, including dividers between left-turn and straight lanes. Conversely, U-Net (middle row of Fig 4) has blurred edges, characterized by noticeable jaggedness along lane dividers, whereas DeepLabv3+ (bottom row of Fig 4) fails to detect some lanes owing to its restricted receptive field. Rural Paths: ISRE (bottom row of Fig 5) effectively preserves details of small paths (e.g., curved sections of field paths) through multi-scale fusion. Conversely, SRCNN (middle row of Fig 5) fails to maintain path continuity due to the lack of contextual information. Skeleton Comparison: The ground-truth skeleton (top row of Fig 5) is nearly precisely aligned with the skeleton of ISRE (bottom row of Fig 5). In contrast, skeletons from other models contain breakpoints or redundant branches (e.g., the U-Net skeleton lacks connectivity at intersections).

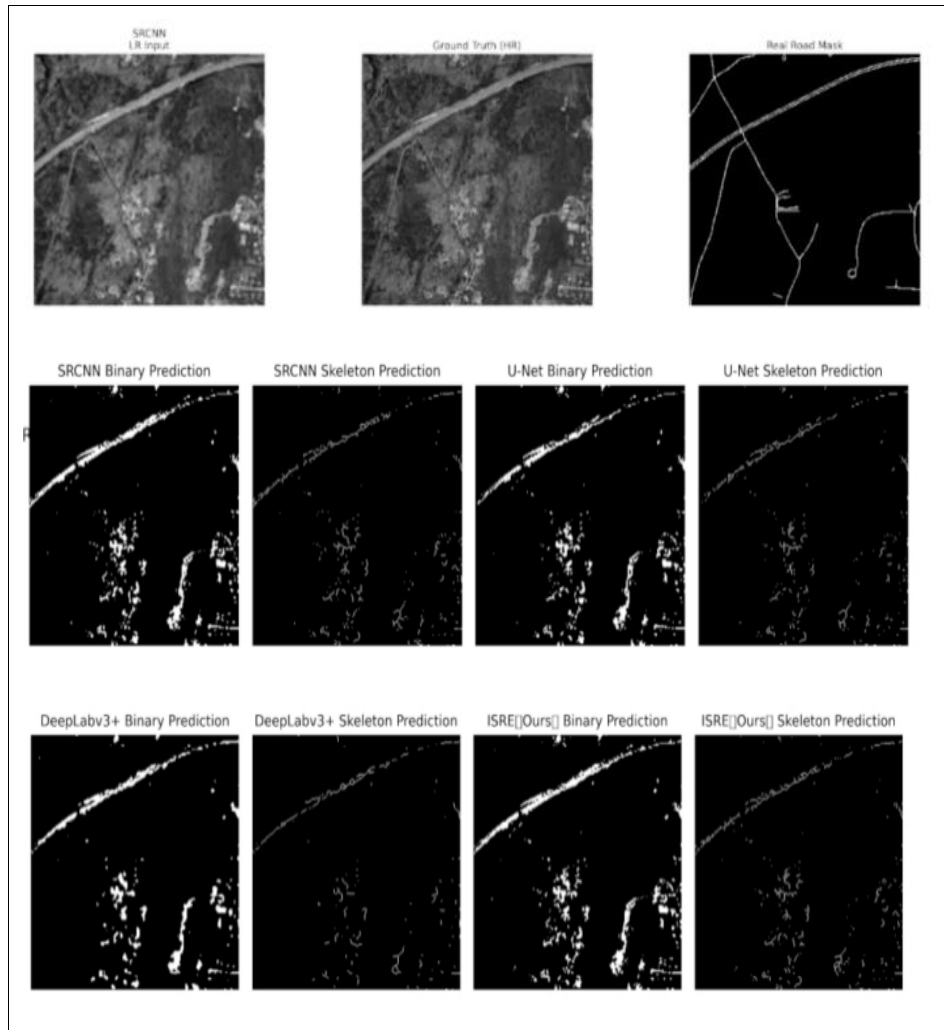


Figure 5: Extraction results of various models (SRCNN, U-Net, DeepLabv3+, ISRE) on rural paths.
Source: Authors, (2026).

Experimental results indicate that ISRE performs optimally in road extraction tasks. The key reasons lie in:

Enhanced Channel Attention (ECA): By dynamically adjusting channel weights, the model focuses more on critical features such as road edges while suppressing background noise interference;

Multi-Scale Feature Fusion (MSF): Fusing convolutional features of different scales compensates for the insufficient representation of complex road structures (e.g., wide and narrow roads) by single-scale features;

Structure-Aware Loss (ASL): Jointly optimizing pixel-level, gradient-level, and region-level losses ensures both overall image similarity and enhanced connectivity and detail preservation of road structures.

Compared with existing methods, the innovative aspect of ISRE lies in integrating structure awareness with multi-scale feature fusion, addressing the issues of incomplete road segmentation and skeleton fragmentation in traditional models under complex scenarios (e.g., intersections and small paths). Future work could further explore lightweight design (e.g., model compression) to adapt to real-time processing requirements on edge devices such as aerial photography platforms.

V. CONCLUSIONS

V.1 CONCLUSION

This article proposes an enhanced SRCNN model called ISRE that integrates structural perception to address the issues of limited feature expression ability, insufficient multi-scale adaptability, and missing structural information in the lightweight SRCNN model for aerial image road extraction tasks. By introducing efficient channel attention module, multi-scale feature fusion module, and adaptive structure aware loss function into the infrastructure, ISRE significantly enhances its ability to focus on key road features, adapt to different scale road shapes, and model road network topology connectivity while maintaining its lightweight advantage. The comprehensive experiment on the Massachusetts road dataset shows that ISRE has achieved significant improvement in F1 score compared to the baseline SRCNN model, especially in the topological accuracy indicator measuring road connectivity, which has improved by more than 11%. At the same time, ISRE achieves more competitive comprehensive performance with much lower parameter count and computational cost than complex segmentation models such as U-Net and DeepLabv3+. This fully validates the effectiveness of the ISRE model design, providing a highly promising solution for efficient and high-precision road extraction tasks on computing resource constrained platforms such as drones and mobile terminals.

V.2 LIMITATIONS AND PROSPECTS

Although ISRE has demonstrated superior performance on public datasets, there are still certain limitations. For example, the performance of the model under extreme lighting conditions, severe occlusion (such as roads under dense vegetation), or special road types that do not appear in the training data (such as temporary construction roads) needs further validation and enhancement. Future research work can be carried out from the following aspects:

- 1) Long distance contextual modeling: Exploring the combination of visual transformers or self-attention mechanisms with existing CNN architectures to more effectively model the long-range dependencies of road networks, thereby further improving extraction accuracy in complex intersections and large road networks.
- 2) Multimodal data fusion: Research on road extraction methods that fuse multiple sources of remote sensing data, such as LiDAR point clouds and hyperspectral images. The precise three-dimensional elevation information provided by LiDAR can effectively assist in distinguishing ground objects (such as squares and parking lots) that are difficult to distinguish in two-dimensional images and have similar spectral features to roads, enhancing the robustness of the model.
- 3) Ultimate optimization of model efficiency: In order to meet the real-time processing requirements of extreme resource constrained scenarios such as satellites and micro drones, future work will explore more aggressive lightweight strategies. This includes but is not limited to: applying neural network architecture search technology to automatically find the optimal combination of lightweight modules; Using knowledge distillation techniques, the structured knowledge of large teacher models (such as high-performance segmentation networks) is transferred to compact student models such as ISRE, in order to approach or even surpass the performance of complex models without significantly increasing computational overhead

VI. AUTHOR'S CONTRIBUTION

Conceptualization: Rongfu Wang, Mary Jane C. Samonte.

Methodology: Rongfu Wang, Mary Jane C. Samonte.

Investigation: Rongfu Wang, Mary Jane C. Samonte..

Discussion of results: Rongfu Wang, Mary Jane C. Samonte.

Writing – Original Draft: Rongfu Wang, Mary Jane C. Samonte.

Writing – Review and Editing: Rongfu Wang, Mary Jane C. Samonte.

Resources: Rongfu Wang, Mary Jane C. Samonte.

Supervision: Rongfu Wang, Mary Jane C. Samonte.

Approval of the final text: Rongfu Wang, Mary Jane C. Samonte.

VII. ACKNOWLEDGMENTS

This work was supported by the Dongguan Science and Technology of Social Development Program (20221800905782).

VIII. REFERENCES

- [1] Y. Alshamaila, S. Papagiannidis, H. Alsawalqah, and I. Aljarah, 'Effective use of smart cities in crisis cases: A systematic review of the literature', *Int. J. Disaster Risk Reduct.*, vol. 85, p. 103521, 2023.
- [2] M. Sadaf et al., 'Connected and automated vehicles: Infrastructure, applications, security, critical challenges, and future aspects', *Technologies*, vol. 11, no. 5, p. 117, 2023.
- [3] W. Eckstein, 'Automatic road extraction based on multi-scale, grouping, and context', *PE RS-....*, 2022.
- [4] H. Ghandorh, W. Boulila, S. Masood, A. Koubaa, F. Ahmed, and J. Ahmad, 'Semantic segmentation and edge detection—Approach to road detection in very high resolution satellite images', *Remote Sens.*, vol. 14, no. 3, p. 613, 2022.
- [5] R. Tamilarasi and S. Prabu, 'Automated building and road classifications from hyperspectral imagery through a fully convolutional network and support vector machine', *J. Supercomput.*, vol. 77, no. 11, pp. 13243–13261, 2021.
- [6] A. Di Benedetto, M. Fiani, and L. M. Gujski, 'U-Net-based CNN architecture for road crack segmentation', *Infrastructures*, vol. 8, no. 5, p. 90, 2023.
- [7] M. N. Mahmud, M. K. Osman, A. P. Ismail, F. Ahmad, K. A. Ahmad, and A. Ibrahim, 'Road image segmentation using unmanned aerial vehicle images and DeepLab V3+ semantic segmentation model', in *2021 11th IEEE International Conference on Control System, Computing and Engineering (ICCSCE)*, IEEE, 2021, pp. 176–181.
- [8] R. Wang and M. J. C. Samonte, 'A Review of Deep Learning in Inverse Problem Imaging Using Super-Resolution Reconstruction Algorithms', *J. Image Graph.*, vol. 12, no. 4, 2024.
- [9] R. Wang and M. J. C. Samonte, 'Research on UAV Image Road Recognition Application Based on Improved SRCNN', in *2024 2nd International Conference on Computer Network Technology and Electronic and Information Engineering (CNTEIE)*, IEEE, 2024, pp. 159–165.
- [10] Y. Peng, D. Z. Chen, and M. Sonka, 'U-net v2: Rethinking the skip connections of u-net for medical image segmentation', in *2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI)*, IEEE, 2025, pp. 1–5.
- [11] L. Jiao et al., 'Multiscale deep learning for detection and recognition: A comprehensive survey', *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 36, no. 4, pp. 5900–5920, 2024.
- [12] Z. Li, H. Li, and L. Meng, 'Model compression for deep neural networks: A survey', *Computers*, vol. 12, no. 3, p. 60, 2023.

- [13] W. Zheng, H. Xu, P. Li, R. Wang, and X. Shao, 'Sac-rsm: A high-performance uav-side road surveillance model based on super-resolution assisted learning', *IEEE Internet Things J.*, 2024.
- [14] C. Peng, P. Shu, X. Huang, Z. Fu, and X. Li, 'LCRCA: image super-resolution using lightweight concatenated residual channel attention networks', *Appl. Intell.*, vol. 52, no. 9, pp. 10045–10059, 2022.
- [15] Y. Liu, Q. Jia, X. Fan, S. Wang, S. Ma, and W. Gao, 'Cross-SRN: Structure-preserving super-resolution network with cross convolution', *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 8, pp. 4927–4939, 2021.
- [16] E. M. Bakr, A. El-Sallab, and M. Rashwan, 'EMCA: Efficient multiscale channel attention module', *IEEE Access*, vol. 10, pp. 103447–103461, 2022.
- [17] D. Shao, L. Ren, and L. Ma, 'MSF-Net: a lightweight multi-scale feature fusion network for skin lesion segmentation', *Biomedicines*, vol. 11, no. 6, p. 1733, 2023.
- [18] C. Ma, Y. Rao, J. Lu, and J. Zhou, 'Structure-preserving image super-resolution', *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 11, pp. 7898–7911, 2021.
- [19] J. Luo, H. Lin, X. Wei, and Y. Wang, 'Adaptive canny and semantic segmentation networks based on feature fusion for road crack detection', *IEEE Access*, vol. 11, pp. 51740–51753, 2023.
- [20] Y. Wei and M. Xu, 'Detection of lane line based on Robert operator', *J. Meas. Eng.*, vol. 9, no. 3, pp. 156–166, 2021.
- [21] J. Liu et al., 'Assessment of flood susceptibility mapping using support vector machine, logistic regression and their ensemble techniques in the Belt and Road region', *Geocarto Int.*, vol. 37, no. 25, pp. 9817–9846, 2022.
- [22] C. Han, T. Ma, G. Xu, S. Chen, and R. Huang, 'Intelligent decision model of road maintenance based on improved weight random forest algorithm', *Int. J. Pavement Eng.*, vol. 23, no. 4, pp. 985–997, 2022.
- [23] D. Pan, M. Zhang, and B. Zhang, 'A generic FCN-based approach for the road-network extraction from VHR remote sensing images—using openstreetmap as benchmarks', *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 14, pp. 2662–2673, 2021.
- [24] R. Azad et al., 'Medical image segmentation review: The success of u-net', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2024.
- [25] Y. Liu, M. Zhu, J. Wang, X. Guo, Y. Yang, and J. Wang, 'Multi-scale deep neural network based on dilated convolution for spacecraft image segmentation', *Sensors*, vol. 22, no. 11, p. 4222, 2022.
- [26] Y. He and L. Xiao, 'Structured pruning for deep convolutional neural networks: A survey', *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 5, pp. 2900–2919, 2023.
- [27] T. Schumacher, M. Strohmaier, and F. Lemmerich, 'A comparative evaluation of quantification methods', *ArXiv Prepr. ArXiv210303223*, 2021.
- [28] G. Habib, S. M. Kaleem, T. Rouf, B. Lall, and others, 'A comprehensive review of knowledge distillation in computer vision', *ArXiv Prepr. ArXiv240400936*, 2024.
- [29] J. Navarro and F. Pellerey, 'Preservation of ILR and IFR aging classes in sums of dependent random variables', *Appl. Stoch. Models Bus. Ind.*, vol. 38, no. 2, pp. 240–261, 2022.
- [30] L. Zhou et al., 'A multi-scale object detector based on coordinate and global information aggregation for UAV aerial images', *Remote Sens.*, vol. 15, no. 14, p. 3468, 2023.
- [31] A. S. M. Miah, Y. S. Hwang, and J. Shin, 'Sensor-based human activity recognition based on multi-stream time-varying features with eca-net dimensionality reduction', *IEEE Access*, 2024.
- [32] Y. Zhang, Q. Zhan, and Z. Ma, 'EfficientNet-ECA: A lightweight network based on efficient channel attention for class-imbalanced welding defects classification', *Adv. Eng. Inform.*, vol. 62, p. 102737, 2024.
- [33] V. Mnih, *Machine learning for aerial image labeling*. University of Toronto (Canada), 2013. Accessed: Aug. 18, 2025. [Online]. Available: <https://search.proquest.com/openview/215c1c8c5424a6facf8885f4ac00c575/1?pq-origsite=gscholar&cbl=18750>
- [34] Y. Wang, L. Yang, X. Liu, and P. Yan, 'An improved semantic segmentation algorithm for high-resolution remote sensing images based on DeepLabv3+', *Sci. Rep.*, vol. 14, no. 1, p. 9716, 2024.
- [35] D. R. I. M. Setiadi, 'PSNR vs SSIM: imperceptibility quality assessment for image steganography', *Multimed. Tools Appl.*, vol. 80, no. 6, pp. 8423–8444, 2021.
- [36] D. J. Hand, P. Christen, and N. Kirielle, 'F*: an interpretable transformation of the F-measure', *Mach. Learn.*, vol. 110, no. 3, pp. 451–456, 2021.
- [37] X. Hu, Y. Wang, L. Fuxin, D. Samaras, and C. Chen, 'Topology-aware segmentation using discrete morse theory', *ArXiv Prepr. ArXiv210309992*, 2021.
- [38] K. Han, Y. Wang, J. Guo, and E. Wu, 'ParameterNet: parameters are all you need for large-scale visual pretraining of mobile networks', in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 15751–15761.