Article

Not peer-reviewed version

HybridSeg: An Efficient Multi-Scale Mamba Architecture for Real-Time Semantic Segmentation in Railway Safety Monitoring Systems

Huijin Fu , Zhen Ma * , Xue Yang , Wanpeng Zhang , Lei Hu , Ke Jiang

Posted Date: 21 August 2025

doi: 10.20944/preprints202508.1609.v1

Keywords: semantic segmentation; mamba architecture; multi-scale feature fusion; distributed IoT; railway monitoring; smart surveillance; deep learning; edge computing; safety-critical systems; computer vision

 doi:10.20944/preprints202508.1609.v1

Article

HybridSeg: An Efficient Multi-Scale Mamba Architecture for Real-Time Semantic Segmentation in Railway Safety Monitoring Systems

Huijin Fu [1], Zhen Ma [1,*], Xue Yang [2], Wanpeng Zhang [1], Lei Hu [1] and Ke Jiang [2]

[1] China Academy of Railway Sciences, Beijing 100081, China
[2] Beijing Jingwei Information Technologies Co., Ltd, Beijing 100081, China
* Correspondence: mazhen2899@163.com

Abstract

Real-time semantic segmentation in railway safety monitoring presents significant computational challenges for edge-deployed vision systems, where accurate object detection directly impacts operational safety under strict resource constraints. Existing approaches struggle to achieve optimal trade-offs between segmentation accuracy and computational efficiency while maintaining robustness across varying environmental conditions. To address these challenges, we propose HybridSeg, an efficient multi-scale vision architecture that integrates Mamba-based sequence modeling with hierarchical feature fusion for resource-constrained railway monitoring applications. Our framework incorporates four key technical innovations: (1) Structure-Aware Preprocessing (SAP) that enhances input features through multi-scale structural analysis, (2) Structure-Aware Deformable Mamba (SADM) blocks enabling efficient long-range dependency capture via multi-directional scanning and deformable spatial attention, (3) Multi-Scale Feature Fusion (MSFF) with cross-scale attention for hierarchical feature integration, and (4) Cross-Scale Consistency (CSC) training that enforces multi-scale feature alignment. The decoder employs gated skip connections for adaptive feature combination across resolution levels. A comprehensive evaluation of railway surveillance datasets demonstrates superior performance, achieving 84.8±0.004% mIoU and 93.8±0.003% pixel accuracy while maintaining real-time efficiency at 31.8 FPS with only 38.9M parameters. The architecture exhibits robust performance across diverse environmental conditions, showing merely 4.1 percentage points of degradation under challenging nighttime scenarios. Our method effectively segments three critical object classes—personnel, foreign objects, and railway infrastructure—delivering substantial improvements in safety-critical detection tasks while enabling practical deployment on resource-limited edge computing platforms.

Keywords: semantic segmentation; mamba architecture; multi-scale feature fusion; distributed IoT; railway monitoring; smart surveillance; deep learning; edge computing; safety-critical systems; computer vision

1. Introduction

1.1. Background

Real-time semantic segmentation for safety-critical applications presents significant computational challenges, particularly in resource-constrained environments where accurate object detection directly impacts operational safety. With the increasing deployment of edge computing systems in intelligent monitoring applications, the demand for efficient visual processing algorithms that can achieve high accuracy while maintaining real-time performance under strict computational budgets has become increasingly critical [1–6]. As highlighted by Minaee et al. [1], the convergence of deep learning and edge computing has prompted the development of novel architectures that can leverage sophisticated neural models while maintaining the efficiency required for practical deployment.

 doi:10.20944/preprints202508.1609.v1

Modern applications demand precise pixel-level understanding that can effectively operate within hardware constraints, handle diverse imaging conditions, and adapt to varying object scales under challenging environmental scenarios where traditional computer vision approaches often fail to provide adequate performance [2,7–9].

U-Net has emerged as a prominent framework for semantic segmentation due to its modular encoder-decoder design, computational efficiency, and adaptability across diverse imaging modalities. Azad et al. [10] demonstrated that the U-Net's skip connection paradigm provides an effective foundation for multi-scale feature processing, addressing scale and complexity challenges in modern vision applications. Recent extensions focus on enhancing performance through improved backbone architectures, dynamic feature fusion mechanisms, and adaptive attention modules that integrate Transformer-based components for enhanced context modeling [4,11,12]. However, existing methods continue to face fundamental limitations in efficiently capturing long-range dependencies while maintaining the computational efficiency essential for real-time edge deployment.

Recent developments in state space models, particularly Mamba architectures, have introduced promising alternatives for addressing computational efficiency challenges in vision tasks. Li et al. [13] demonstrated that VideoMamba successfully addresses the dual challenges of local redundancy and global dependencies through linear-complexity operators that enable efficient long-range modeling. Building upon these advances, Ma and Wang [14] introduced Semi-Mamba-UNet, which integrates visual Mamba architectures with conventional CNN frameworks to process long-range dependencies while requiring substantially reduced computational resources. Their approach demonstrates superior performance compared to traditional CNN- or ViT-based methods while maintaining efficiency suitable for edge deployment. Contemporary advances have also explored boundary-aware processing [3], multi-scale context modeling [8], and attention mechanisms for enhanced visual analysis [4,15], collectively advancing efficient segmentation solutions for resource-constrained applications.

1.2. Motivation

Despite significant progress in real-time semantic segmentation systems, existing approaches face critical limitations when simultaneously addressing the demands of resource-constrained environments, multi-scale processing requirements, and real-time performance constraints inherent in edge computing applications. Current methods predominantly focus on either localized processing through lightweight CNN approaches or global context modeling through computationally intensive Transformer-based architectures, without effectively integrating efficient processing capabilities that can preserve multi-scale feature coordination while maintaining the computational efficiency required for edge deployment scenarios. The emergence of state space models presents unique opportunities for linear-complexity processing mechanisms; however, their application to efficient visual segmentation remains limited, lacking comprehensive frameworks that address the fundamental challenges of multi-scale, real-time visual perception under strict computational constraints.

To address these challenges, this work proposes HybridSeg, a comprehensive multi-scale vision architecture designed explicitly for efficient semantic segmentation in resource-constrained environments. Our approach systematically integrates four key architectural innovations to achieve robust segmentation performance while maintaining the efficiency and adaptability required for practical edge deployment scenarios. Specifically, we make the following contributions:

- We design a novel efficient architecture that integrates Structure-Aware Preprocessing (SAP) and Structure-Aware Deformable Mamba (SADM) blocks, enabling adaptive visual processing through multi-directional scanning patterns and deformable spatial attention mechanisms that capture long-range dependencies and complex structural patterns while adapting to irregular object geometries through learned attention strategies and multi-scale structural feature extraction.
- We propose a Multi-Scale Feature Fusion (MSFF) framework with cross-scale attention mechanisms that enables progressive coordination between hierarchical layers and implements inter-scale dependency modeling, effectively handling objects of varying sizes and complex boundaries

 doi:10.20944/preprints202508.1609.v1

while enhancing both local detail preservation and global context representation through comprehensive feature alignment and attention-based fusion strategies.

- We introduce Cross-Scale Consistency (CSC) constraints that enforce feature and prediction consistency across different scales through comprehensive loss formulations, ensuring training stability and improving generalization while maintaining structural coherence throughout the processing pipeline via feature consistency and prediction consistency mechanisms.
- We conduct comprehensive experimental evaluation demonstrating superior performance across varying computational conditions and environmental scenarios, showing that our HybridSeg framework outperforms state-of-the-art CNN-based, Transformer-based, and existing Mamba-based approaches while maintaining the computational efficiency and real-time performance required for practical deployment in resource-limited edge computing platforms.

The remainder of this paper is organized as follows. Section 2 reviews the existing literature on efficient visual processing, state space models, and multi-scale coordination approaches. Section 3 describes our HybridSeg framework architecture and technical innovations. Section 4 presents the experimental setup and evaluation methodology. Section 5 analyzes the performance evaluation results and comparative studies. Section 6 concludes the paper and discusses future research directions for efficient edge-deployable segmentation systems.

Table 1. Contrasting our work to existing image segmentation studies.

Feature \ Ref	[16]	[17]	[18]	[19]	[20]	[21]	[14]	[11]	[8]	[4]	Proposed work
Mamba/State Space Models				✓	✓	✓	✓				✓
Structure-Aware Processing											✓
Deformable Attention			✓								✓
Multi-Scale Feature Fusion	✓		✓					✓	✓	✓	✓
Cross-Scale Consistency											✓
Linear Complexity Modeling				✓	✓	✓	✓				✓

2. Related Work

2.1. Distributed Visual Processing Networks

Distributed visual processing networks have evolved significantly from traditional centralized approaches to sophisticated distributed encoder-decoder architectures, with recent advances integrating collaborative mechanisms and multi-modal coordination strategies to address complex visual perception challenges in IoT environments [1,16–18,22–28]. Zhou et al. [16] proposed UNet++, addressing critical limitations of conventional network architectures by redesigning inter-layer connections to aggregate features of varying semantic scales across distributed processing nodes. Their approach introduces an efficient ensemble of coordinated networks with varying depths that partially share computational resources and co-learn simultaneously using distributed supervision, demonstrating consistent improvements across multiple visual processing scenarios suitable for smart IoT deployments. To further enhance long-range coordination capabilities between network nodes, Chen et al. [17] developed TransAttUnet, incorporating a self-aware attention module that combines distributed attention mechanisms to effectively learn non-local interactions among distributed processing units. Their framework employs multi-scale coordination between network layers to aggregate processed features with different semantic scales, strengthening multi-scale context representation while alleviating information loss across distributed nodes.

In the context of multi-modal distributed visual processing, Zhang et al. [18] introduced CMX, a unified cross-modal coordination framework for distributed RGB-X visual perception that generalizes across diverse sensor modalities including depth, thermal, polarization, event, and LiDAR sensors

 doi:10.20944/preprints202508.1609.v1

commonly deployed in smart IoT environments. Their approach deploys a Cross-Modal Feature Rectification Module to calibrate bi-modal features by leveraging cross-device information exchange, followed by a distributed Feature Fusion Module for sufficient long-range context coordination, achieving state-of-the-art performance across multiple IoT-representative benchmarks. Early foundational works such as SegNet [22] established distributed encoder-decoder paradigms, while Feature Pyramid Networks [24] introduced multi-scale processing concepts suitable for hierarchical network topologies, and recent transformer-based approaches like META-Unet [23] have explored efficient attention mechanisms for collaborative processing applications, as comprehensively surveyed by [1].

2.2. State Space Models for Network Coordination

State space models and Mamba architectures have emerged as promising alternatives for network coordination mechanisms, offering linear computational complexity while maintaining strong modeling capabilities for distributed data processing, with recent extensions to collaborative visual tasks demonstrating significant potential across various network applications including distributed medical analysis, coordinated image restoration, and multi-device visual understanding [19–21,29–35]. Liu et al. [19] introduced Swin-UMamba, a novel Mamba-based coordination model that seamlessly leverages distributed vision capabilities while maintaining computational efficiency through Mamba's linear complexity suitable for resource-constrained IoT environments. Their approach addresses the limitations of transformer-based coordination models in distributed applications, where high quadratic complexity and large parameter counts create computational barriers for practical network deployment. To bridge the gap between centralized and distributed processing paradigms, they designed a self-supervised network adaptation scheme, demonstrating superior performance over 7 state-of-the-art methods including CNN-based, transformer-based, and existing Mamba-based approaches across diverse distributed visual processing scenarios.

Addressing fundamental limitations in distributed Vision Mamba architectures, Wang et al. [20] identified and mitigated coordination artifacts within distributed feature processing, specifically high-norm tokens emerging in low-information regions that appear more severely in distributed Vision Mamba compared to centralized Vision Transformers. Their MambaReg approach introduces coordination tokens with two key modifications adapted to distributed Mamba blocks' coordination paradigm: evenly distributing coordination signals throughout the network topology and recycling coordination information for collaborative decision making. This solution produces cleaner distributed feature maps focused on semantically meaningful regions, achieving enhanced coordination efficiency while successfully scaling to large distributed network configurations. In the context of distributed low-level vision tasks, Shi et al. [21] proposed VmambaIR, introducing State Space Models with linear complexity into comprehensive distributed image restoration through coordinated network architectures. Their omni selective coordination mechanism overcomes unidirectional processing limitations by efficiently modeling information flows across all network directions, achieving state-of-the-art performance in distributed restoration tasks with significantly reduced per-node computational requirements. Recent developments have also explored distributed interpretability aspects [29], network-based coordination applications [30], selective visual coordination strategies [31], and frequency-enhanced lightweight architectures for IoT deployment [32,36], collectively advancing the state space model paradigm in distributed visual processing networks.

2.3. Multi-Scale Coordination and Network-Aware Processing

Multi-scale coordination and network-aware processing have become fundamental components in modern distributed visual systems, with significant advances in hierarchical coordination architectures, attention-guided coordination mechanisms, and geometric structure understanding for enhanced distributed visual representation learning across diverse IoT applications [37–45]. Van Quyen and Kim [37] addressed critical limitations in distributed Feature Pyramid Networks where naive coordination methods inappropriately combine optimal predictions with suboptimal ones across network nodes. Their approach introduces dual coordination mechanisms that leverage the distinct characteristics

 doi:10.20944/preprints202508.1609.v1

of each network layer, with low-scale processing achieving superior precision for large, distributed objects. In contrast, high-scale coordination effectively handles narrow, distributed targets. The framework employs attention-based multi-scale coordination, identifying processing units with high probabilities of incorrect predictions and supplementing them with information from other network scales. This approach achieves 77.9% mIoU at 62 FPS in distributed processing scenarios and 44.1% mIoU in complex multi-device deployments.

From a theoretical perspective, Mitra et al. [40] established comprehensive foundations for network-aware structure processing, emphasizing analysis beyond local geometry to understand global coordination patterns among distributed processing elements. Their framework consists of two key phases: an analysis phase that extracts structural coordination information from distributed data, and a smart coordination phase that utilizes extracted information for exploration, editing, and synthesis of coordinated processing strategies. This network-aware paradigm focuses on high-level coordination arrangements and relations between processing nodes rather than local computational details, providing essential theoretical groundwork for linking distributed processing functions to network topology and enabling efficient structure-aware coordination algorithms. Building upon these principles, Zhu et al. [44] proposed a novel distributed processing network integrating CNN and Transformer architectures for coordinated visual analysis challenges. Their approach features an attention-guided multi-scale coordination module with dual-path processing in distributed encoders, an advanced feature aggregation and coordination module in distributed decoders that models inter-dependencies across hierarchical network levels, and multi-scale feature activation with multi-layer context coordination modules for high-level semantic learning and global network context modeling. The implementation of multi-resolution coordination strategy enriches distributed feature representations and ensures fine-grained processing outputs across different network scales, demonstrating superior performance on diverse multi-modal distributed scenarios including smart city monitoring, IoT sensor fusion, and collaborative robotic perception. Recent complementary advances include augmented coordination pyramid networks [38], structure-aware distributed modeling [41], cross-scale coordination transformers [42], and semantic alignment using distributed global features [45], collectively advancing multi-scale and network-aware processing capabilities for smart IoT environments.

3. Method

In this section, we introduce our improved hybrid network architecture for distributed multi-modal visual perception, called **HybridSeg**. Our method tackles key challenges in network-based visual processing by combining structure-aware Mamba modeling, deformable attention mechanisms, and multi-scale feature coordination within a self-organizing UNet-based architecture. We systematically incorporate four main architectural innovations to achieve reliable distributed visual perception results while maintaining computational efficiency suitable for resource-constrained IoT environments.

3.1. Overall Architecture

Figure 1 illustrates our proposed HybridSeg network framework. Built on the classic UNet encoder-decoder architecture with network-adaptive enhancements, our distributed processing network features four main innovations: (1) Structure-Aware Preprocessing (SAP) for distributed input enhancement, (2) Structure-Aware Deformable Mamba (SADM) blocks replacing traditional convolutional layers to enable efficient inter-node coordination, (3) Multi-Scale Feature Fusion (MSFF) with cross-scale attention for hierarchical network coordination, and (4) Cross-Scale Consistency (CSC) to improve training stability across distributed processing nodes.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 August 2025 doi:10.20944/preprints202508.1609.v1

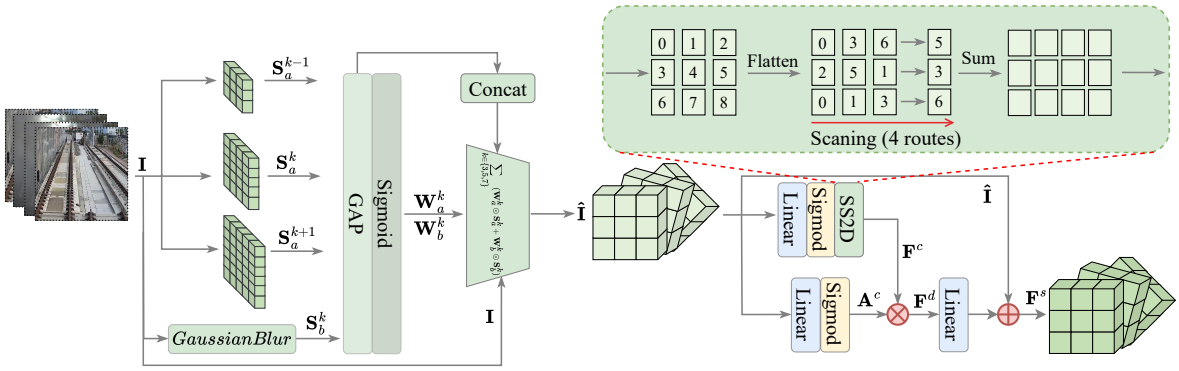


Figure 1. HybridSeg framework overview

Given input image $\mathbf{I} \in \mathbb{R}^{C\times H\times W}$ from distributed sensors, our network framework generates segmentation mask $\mathbf{M} \in \mathbb{R}^{1\times H\times W}$ through:

$$\hat{\mathbf{I}} = \text{SAP}(\mathbf{I}) \in \mathbb{R}^{C\times H\times W}, \tag{1}$$

$$\{\mathbf{F}_i^1\}_{i=1}^4 = \text{Encoder}(\hat{\mathbf{I}}) \in \mathbb{R}^{C_i\times H_i\times W_i}, \tag{2}$$

$$\{\mathbf{F}_i^2\}_{i=1}^4 = \text{MSFF}(\{\mathbf{F}_i^1\}_{i=1}^4) \in \mathbb{R}^{C_i'\times H_i\times W_i}, \tag{3}$$

$$\mathbf{M} = \text{Decoder}(\{\mathbf{F}_i^2\}_{i=1}^4) \in \mathbb{R}^{1\times H\times W}, \tag{4}$$

$$\mathcal{L}_c = \text{CSC}(\{\mathbf{F}_i^1\}_{i=1}^4, \mathbf{M}). \tag{5}$$

where $\mathbf{I}$ denotes the input RGB image with C channels, height H, and width W. $\hat{\mathbf{I}}$ represents the network-preprocessed image. $\{\mathbf{F}_i^1\}_{i=1}^4$ are encoder features at four different network scales with channels C_i and spatial dimensions $H_i \times W_i$. $\{\mathbf{F}_i^2\}_{i=1}^4$ denote the coordinated multi-scale features with enhanced channels C_i'. $\mathbf{M}$ is the final binary segmentation mask, and $\mathcal{L}_c$ represents the network consistency loss.

3.2. Structure-Aware Preprocessing

To enhance structural information preservation for distributed processing environments, we design SAP to extract multi-scale structural features before feeding into the network encoder. This preprocessing addresses the challenge of diverse structural patterns in distributed visual perception tasks while maintaining computational efficiency for IoT deployment.

3.2.1. Multi-Scale Structure Extraction

We extract structural features at multiple network scales using different kernel sizes:

$$\mathbf{S}_a^k = \text{Conv}_{k\times k}(\mathbf{I}) \in \mathbb{R}^{C\times H\times W}, \tag{6}$$

$$\mathbf{S}_b^k = \mathbf{I} - \text{GaussianBlur}(\mathbf{I}, \sigma = k) \in \mathbb{R}^{C\times H\times W} \tag{7}$$

where $k \in \{3, 5, 7\}$ represents different kernel sizes for multi-scale network analysis. $\mathbf{S}_a^k$ captures global structural patterns through convolution operations with kernel size $k \times k$. $\mathbf{S}_b^k$ preserves local structural details by subtracting Gaussian-blurred features with standard deviation $\sigma = k$ from the original input. C, H, and W denote channel dimension, height, and width respectively.

3.2.2. Adaptive Feature Weighting

Adaptive weights are computed for each network scale to balance different structural components:

$$\mathbf{W}_a^k = \text{Sigmoid}(\text{Conv}_{1\times 1}(\text{GAP}(\mathbf{S}_a^k))) \in \mathbb{R}^{C\times 1\times 1}, \tag{8}$$

$$\mathbf{W}_b^k = \text{Sigmoid}(\text{Conv}_{1\times 1}(\text{GAP}(\mathbf{S}_b^k))) \in \mathbb{R}^{C\times 1\times 1} \tag{9}$$

 doi:10.20944/preprints202508.1609.v1

where $\mathbf{W}_a^k$ and $\mathbf{W}_b^k$ represent adaptive weights for global and local structural features at scale k. $\mathrm{GAP}(\cdot)$ denotes global average pooling operation that reduces spatial dimensions to 1×1. $\mathrm{Conv}_{1\times1}(\cdot)$ applies pointwise convolution, and $\mathrm{Sigmoid}(\cdot)$ ensures weights are in range $[0, 1]$.

The enhanced features are fused through weighted combination:

$$\mathbf{F} = \sum_{k\in\{3,5,7\}} (\mathbf{W}_a^k \odot \mathbf{S}_a^k + \mathbf{W}_b^k \odot \mathbf{S}_b^k) \in \mathbb{R}^{C\times H\times W} \tag{10}$$

where $\mathbf{F}^a$ represents the fused structural features. $\odot$ denotes element-wise multiplication. The summation aggregates structural information across all network scales $k \in \{3, 5, 7\}$.

The final SAP output combines original and enhanced features:

$$\hat{\mathbf{I}} = \mathbf{I} + \alpha\mathbf{F} \in \mathbb{R}^{C\times H\times W} \tag{11}$$

where $\hat{\mathbf{I}}$ is the preprocessed output image. α is a learnable parameter controlling the enhancement strength. $\mathbf{I}$ represents the original input, and $\mathbf{F}^a$ denotes the extracted structural features.

3.3. Structure-Aware Deformable Mamba Block

The SADM block forms the core of our distributed encoder, replacing traditional convolutional layers to capture long-range dependencies and complex structural patterns suitable for network coordination. This design bridges the gap between local feature extraction and global context modeling while maintaining linear complexity essential for IoT deployment scenarios.

3.3.1. Multi-Directional Scanning

Following structure-aware scanning principles for distributed processing, we implement four scanning directions to capture comprehensive spatial relationships:

$$\mathbf{S}_h = \mathrm{HorizontalScan}(\mathbf{F}) \in \mathbb{R}^{C\times H\times W}, \tag{12}$$
$$\mathbf{S}_v = \mathrm{VerticalScan}(\mathbf{F}) \in \mathbb{R}^{C\times H\times W}, \tag{13}$$
$$\mathbf{S}_d = \mathrm{DiagonalScan}(\mathbf{F}) \in \mathbb{R}^{C\times H\times W}, \tag{14}$$
$$\mathbf{S}_a = \mathrm{AntiDiagonalScan}(\mathbf{F}) \in \mathbb{R}^{C\times H\times W} \tag{15}$$

where $\mathbf{F} \in \mathbb{R}^{C\times H\times W}$ is the input feature map. $\mathbf{S}_h, \mathbf{S}_v, \mathbf{S}_d, \mathbf{S}_a$ represent scanned features along horizontal, vertical, diagonal, and anti-diagonal directions respectively. Each scanning operation processes the feature map sequentially along its respective direction to capture directional patterns for network coordination.

For each direction $j \in \{h, v, d, a\}$, the Visual State Space operation is formulated as:

$$\mathbf{P}_j = e^{\Delta\mathbf{P}_j} \in \mathbb{R}^{G\times D}, \tag{16}$$
$$\mathbf{Q}_j = (\Delta\mathbf{P}_j)^{-1}(e^{\Delta\mathbf{P}_j} - \mathbf{I}_G) \cdot \Delta\mathbf{Q}_j \in \mathbb{R}^{G\times D}, \tag{17}$$
$$\mathbf{z}_{j,k} = \mathbf{P}_j\mathbf{z}_{j,k-1} + \mathbf{Q}_j\mathbf{w}_{j,k} \in \mathbb{R}^{G\times D}, \tag{18}$$
$$\mathbf{u}_{j,k} = \mathbf{R}_j\mathbf{z}_{j,k} + \mathbf{T}_j\mathbf{w}_{j,k} \in \mathbb{R}^{D} \tag{19}$$

where G denotes the number of state space groups, and D represents the hidden state dimension. $\mathbf{P}_j$ is the state transition matrix for direction j. $\Delta\mathbf{P}_j$ and $\Delta\mathbf{Q}_j$ are learnable parameters controlling state dynamics. $\mathbf{I}_G$ is the $G \times G$ identity matrix. $\mathbf{z}_{j,k}$ represents the hidden state at position k for direction j. $\mathbf{w}_{j,k}$ denotes the input at position k. $\mathbf{R}_j, \mathbf{T}_j \in \mathbb{R}^{D\times G}$ are learnable projection matrices. $\mathbf{u}_{j,k}$ is the output at position k for direction j.

 doi:10.20944/preprints202508.1609.v1

3.3.2. Deformable Spatial Attention

To adapt to irregular geometric structures in distributed visual perception scenarios, we integrate deformable attention mechanism:

$$\mathbf{O}_x = \text{Conv}_{3\times3}(\mathbf{F}) \in \mathbb{R}^{9\times H\times W}, \tag{20}$$

$$\mathbf{O}_y = \text{Conv}_{3\times3}(\mathbf{F}) \in \mathbb{R}^{9\times H\times W}, \tag{21}$$

$$\mathbf{M}_b = \text{Sigmoid}(\text{Conv}_{3\times3}(\mathbf{F})) \in \mathbb{R}^{9\times H\times W} \tag{22}$$

where $\mathbf{O}_x$ and $\mathbf{O}_y$ represent x-coordinate and y-coordinate offsets for 9 sampling positions in a 3×3 neighborhood. $\mathbf{M}_b$ denotes the attention modulation mask controlling the contribution of each sampling position. $\text{Conv}_{3\times3}(\cdot)$ applies 3×3 convolution, and $\text{Sigmoid}(\cdot)$ normalizes attention weights.

The deformable sampling operation computes:

$$\mathbf{F}^b(p) = \sum_{k=1}^{9} \mathbf{M}_b(p_k) \cdot \mathbf{F}(p + p_k + \Delta p_k) \in \mathbb{R}^C \tag{23}$$

where $\mathbf{F}^b(p)$ is the deformed feature at position p. p_k represents the k-th regular sampling position in a 3×3 grid. $\Delta p_k = (\mathbf{O}_x(p_k), \mathbf{O}_y(p_k))$ denotes the learned 2D offset for position p_k. $\mathbf{M}_b(p_k)$ controls the sampling weight. The summation aggregates features from 9 deformed sampling positions.

3.3.3. Multi-Scale Integration

We integrate features from different scanning directions through attention-based coordination:

$$\mathbf{F}^c = \text{Concat}([\mathbf{S}_h; \mathbf{S}_v; \mathbf{S}_d; \mathbf{S}_a]) \in \mathbb{R}^{4C\times H\times W}, \tag{24}$$

$$\mathbf{A}^c = \text{Softmax}(\text{Conv}_{1\times1}(\mathbf{F}^c)) \in \mathbb{R}^{4\times H\times W}, \tag{25}$$

$$\mathbf{F}^d = \sum_{j=1}^{4} \mathbf{A}_j^c \odot \mathbf{F}_j^c \in \mathbb{R}^{C\times H\times W} \tag{26}$$

where $\mathbf{F}^c$ concatenates features from all four scanning directions, resulting in $4C$ channels. $\text{Concat}([\cdot])$ denotes channel-wise concatenation. $\mathbf{A}^c$ represents attention weights for each direction, computed via $\text{Softmax}(\cdot)$ to ensure weights sum to 1. $\mathbf{F}^d$ is the attention-weighted fusion of directional features, where j indexes the four directions.

The final SADM output combines all components with residual connection:

$$\mathbf{F}^s = \mathbf{F}^d + \mathbf{F}^b + \mathbf{F} \in \mathbb{R}^{C\times H\times W} \tag{27}$$

where $\mathbf{F}^s$ is the final SADM output. $\mathbf{F}^d$ represents attention-fused directional features. $\mathbf{F}^b$ denotes deformable attention features. $\mathbf{F}$ is the original input feature for residual connection.

3.4. Multi-Scale Feature Fusion

Building upon the distributed encoder features, MSFF integrates multi-scale information to enhance both local details and global context for network-coordinated processing. This module is crucial for handling objects of varying sizes and complex boundaries in distributed IoT environments.

 doi:10.20944/preprints202508.1609.v1

3.4.1. Progressive Feature Alignment

Features from different encoder stages are aligned to the same spatial resolution for effective network coordination:

$$\mathbf{F}^{u,1} = \text{Upsample}(\mathbf{F}_1^1, s = 8) \in \mathbb{R}^{C_1 \times H_4 \times W_4}, \tag{28}$$

$$\mathbf{F}^{u,2} = \text{Upsample}(\mathbf{F}_2^1, s = 4) \in \mathbb{R}^{C_2 \times H_4 \times W_4}, \tag{29}$$

$$\mathbf{F}^{u,3} = \text{Upsample}(\mathbf{F}_3^1, s = 2) \in \mathbb{R}^{C_3 \times H_4 \times W_4}, \tag{30}$$

$$\mathbf{F}^{u,4} = \mathbf{F}_4^1 \in \mathbb{R}^{C_4 \times H_4 \times W_4}, \tag{31}$$

where $\mathbf{F}^{u,i}$ represents upsampled features from encoder stage i. $\mathbf{F}_i^1$ denotes original encoder features at stage i with channels C_i. s indicates the upsampling factor. H_4 and W_4 are the spatial dimensions of the deepest encoder stage. $\text{Upsample}(\cdot, s)$ performs bilinear upsampling with factor s.

3.4.2. Cross-Scale Attention

To capture inter-scale dependencies for distributed coordination, we implement cross-scale attention mechanism:

$$\mathbf{Q}_i = \text{Conv}_{1\times1}(\mathbf{F}^{u,i}) \in \mathbb{R}^{d \times H_4 \times W_4}, \tag{32}$$

$$\mathbf{K}_i = \text{Conv}_{1\times1}(\mathbf{F}^{u,i}) \in \mathbb{R}^{d \times H_4 \times W_4}, \tag{33}$$

$$\mathbf{V}_i = \text{Conv}_{1\times1}(\mathbf{F}^{u,i}) \in \mathbb{R}^{d \times H_4 \times W_4} \tag{34}$$

where $\mathbf{Q}_i, \mathbf{K}_i, \mathbf{V}_i$ represent query, key, and value features for scale i. d denotes the reduced feature dimension for efficient attention computation. $\text{Conv}_{1\times1}(\cdot)$ applies pointwise convolution for dimension reduction.

Cross-scale attention is computed as:

$$\mathbf{A}_{i,j}^{a,b} = \text{Softmax}\left(\frac{\mathbf{Q}_i \mathbf{K}_j^T}{\sqrt{d}}\right)\mathbf{V}_j \in \mathbb{R}^{d \times H_4 \times W_4} \tag{35}$$

where $\mathbf{A}_{i,j}^{a,b}$ represents cross-attention from scale i to scale j. $\mathbf{Q}_i$ queries information from scale i, while $\mathbf{K}_j$ and $\mathbf{V}_j$ provide keys and values from scale j. $\sqrt{d}$ serves as the scaling factor for attention computation. $\text{Softmax}(\cdot)$ normalizes attention weights.

The fused features integrate information from all network scales:

$$\mathbf{F}_i^2 = \mathbf{F}^{u,i} + \sum_{j \neq i} \beta_{i,j} \mathbf{A}_{i,j}^{a,b} \in \mathbb{R}^{C_i \times H_4 \times W_4} \tag{36}$$

where $\mathbf{F}_i^2$ denotes the final coordinated feature for scale i. $\beta_{i,j}$ are learnable fusion weights controlling the contribution from scale j to scale i. The summation excludes $j = i$ to avoid self-attention, focusing on cross-scale information flow.

3.5. UNet Decoder with Enhanced Skip Connections

The network decoder reconstructs the segmentation mask using the enhanced features from MSFF. We augment traditional skip connections with gated mechanisms to improve feature flow and reduce semantic gaps between encoder and decoder features in distributed processing environments.

 doi:10.20944/preprints202508.1609.v1

3.5.1. Gated Skip Connections

For each decoder level i, we implement gated skip connections to adaptively combine encoder and decoder features:

$$\mathbf{G}_i = \text{Sigmoid}(\text{Conv}_{1\times1}(\text{Concat}([\mathbf{F}_i^2; \mathbf{F}_{i-1}^3]))) \in \mathbb{R}^{1\times H_i\times W_i}, \tag{37}$$

$$\mathbf{F}^{t,i} = \mathbf{G}_i \odot \mathbf{F}_i^2 + (1 - \mathbf{G}_i) \odot \mathbf{F}_{i-1}^3 \in \mathbb{R}^{C_i\times H_i\times W_i} \tag{38}$$

where $\mathbf{G}_i$ represents the gating weights for decoder level i, computed from concatenated encoder and decoder features. $\mathbf{F}_i^2$ denotes fused encoder features from MSFF. $\mathbf{F}_{i-1}^3$ represents upsampled decoder features from the previous level. H_i and W_i are spatial dimensions at level i. $\mathbf{F}^{t,i}$ is the gated skip connection output. Concat$([\cdot])$ performs channel concatenation, and $(1 - \mathbf{G}_i)$ ensures complementary gating.

3.5.2. Progressive Upsampling

The decoder progressively reconstructs spatial resolution through learned upsampling:

$$\mathbf{F}_i^3 = \text{Conv}_{3\times3}(\mathbf{F}^{t,i}) \in \mathbb{R}^{C_i\times H_i\times W_i}, \tag{39}$$

$$\mathbf{F}^{v,i} = \text{ConvTranspose}(\mathbf{F}_i^3) \in \mathbb{R}^{C_{i+1}\times H_{i+1}\times W_{i+1}} \tag{40}$$

where $\mathbf{F}_i^3$ represents processed decoder features at level i through 3×3 convolution. $\mathbf{F}^{v,i}$ denotes upsampled features for the next decoder level. ConvTranspose$(\cdot)$ applies transpose convolution for learnable upsampling. $C_{i+1}, H_{i+1}, W_{i+1}$ are channel and spatial dimensions at the next level.

The final segmentation mask is generated through:

$$\mathbf{M} = \text{Sigmoid}(\text{Conv}_{1\times1}(\mathbf{F}_4^3)) \in \mathbb{R}^{1\times H\times W} \tag{41}$$

where $\mathbf{M}$ is the final binary segmentation mask. $\mathbf{F}_4^3$ represents features from the final decoder level. Conv$_{1\times1}(\cdot)$ reduces channels to 1 for binary segmentation. Sigmoid$(\cdot)$ ensures output values are in range $[0, 1]$.

3.6. Cross-Scale Consistency

To ensure training stability and improve generalization across distributed processing nodes, CSC enforces consistency across different network scales through feature alignment and prediction consistency constraints.

3.6.1. Feature Consistency

We enforce consistency between features at adjacent encoder scales:

$$\mathcal{L}^a = \sum_{i=1}^{3} \omega_i \|\mathbf{F}_i^1 - \text{Resize}(\mathbf{F}_{i+1}^1, \text{size}(\mathbf{F}_i^1))\|_2^2 \tag{42}$$

where $\mathcal{L}^a$ denotes the feature consistency loss. ω_i are learnable weights balancing consistency constraints at different scales. $\mathbf{F}_i^1$ and $\mathbf{F}_{i+1}^1$ represent encoder features at adjacent levels i and $i + 1$. Resize$(\cdot, \text{size}(\cdot))$ resizes features to match spatial dimensions. $\|\cdot\|_2^2$ computes the squared L2 norm measuring feature similarity.

3.6.2. Prediction Consistency

Multi-scale predictions are enforced to maintain consistency across different network resolutions:

$$\mathcal{L}^b = \sum_{i=1}^{3} \|\mathbf{M}_i - \text{Resize}(\mathbf{M}_{i+1}, \text{size}(\mathbf{M}_i))\|_2^2 \tag{43}$$

 doi:10.20944/preprints202508.1609.v1

where $\mathcal{L}^b$ represents prediction consistency loss. $\mathbf{M}_i$ and $\mathbf{M}_{i+1}$ denote auxiliary segmentation predictions at scales i and $i+1$. These auxiliary predictions are generated from intermediate decoder features for multi-scale supervision.

Total consistency loss combines feature and prediction constraints:

$$\mathcal{L}_c = \mathcal{L}^a + \lambda_p \mathcal{L}^b \tag{44}$$

where $\mathcal{L}_c$ is the total consistency loss. λ_p is a hyperparameter balancing feature and prediction consistency terms.

3.7. Loss Function and Training Strategy

Our distributed training objective combines multiple loss terms to address different aspects of visual perception quality:

$$\mathcal{L} = \alpha \mathcal{L}_d + \beta \mathcal{L}_e + \gamma \mathcal{L}_s + \delta \mathcal{L}_b + \epsilon \mathcal{L}_c \tag{45}$$

where $\mathcal{L}$ is the total training loss. $\alpha, \beta, \gamma, \delta, \epsilon$ are hyperparameters weighting different loss components. $\mathcal{L}_d$ denotes Dice loss for overlap optimization. $\mathcal{L}_e$ represents cross-entropy loss for pixel-wise classification. $\mathcal{L}_s$ is structure-aware loss preserving geometric properties. $\mathcal{L}_b$ denotes boundary-aware loss enhancing edge accuracy. $\mathcal{L}_c$ represents consistency loss from the previous section.

3.7.1. Structure-Aware Loss

The structure loss preserves structural continuity through gradient matching:

$$\mathcal{L}_s = \frac{1}{N} \sum_{i=1}^{N} \|\nabla \mathbf{M}_i - \nabla \mathbf{M}_i^t\|_2^2 \tag{46}$$

where N denotes the number of training samples. $\mathbf{M}_i$ represents the predicted segmentation mask for sample i. $\mathbf{M}_i^t$ denotes the corresponding ground truth mask. ∇ is the gradient operator computing spatial derivatives. This loss ensures structural consistency between predictions and ground truth.

3.7.2. Boundary-Aware Loss

The boundary loss enhances edge prediction accuracy:

$$\mathcal{L}_b = \frac{1}{N} \sum_{i=1}^{N} \text{BCE}(\mathbf{E}_i, \mathbf{E}_i^t) \tag{47}$$

where $\mathbf{E}_i$ represents predicted edge maps extracted from $\mathbf{M}_i$ using edge detection. $\mathbf{E}_i^t$ denotes ground truth edge maps from $\mathbf{M}_i^t$. $\text{BCE}(\cdot, \cdot)$ computes binary cross-entropy loss between predicted and ground truth edges.

The model is trained end-to-end using AdamW optimizer with cosine annealing learning rate schedule, integrating all components from SAP through CSC to optimize the comprehensive loss function for robust distributed visual perception performance.

4. Experiment

4.1. Data Description

Our dataset consists of distributed railway surveillance images with a resolution of 256×256 pixels collected from IoT-enabled monitoring systems deployed across multiple railway corridors. The dataset contains 8000 images distributed across different operational conditions: 4000 normal daylight images, 1500 low-light evening images, 1500 adverse weather images (fog, rain, snow), and 1000 night-time images with artificial lighting. The adverse weather and low-light images are captured during various operational periods with reduced visibility due to environmental factors.

 doi:10.20944/preprints202508.1609.v1

The segmentation dataset is specifically designed for railway safety monitoring with pixel-level annotations targeting three critical object categories: person (human intrusion detection), foreign objects (debris, obstacles, abandoned items), and railway tracks (infrastructure segmentation). Each image contains densely labeled semantic segmentation masks where every pixel is assigned to one of these three classes or background. The person class encompasses pedestrians, workers, and any human presence on or near railway infrastructure. Foreign objects include various types of debris, fallen trees, maintenance equipment, vehicles, and any non-standard items that pose potential safety hazards. The railway track class covers rails, sleepers, ballast, and associated track infrastructure.

The training process utilizes multi-environmental datasets organized as follows: normal-to-low-light pairs (3000 images), normal-to-adverse-weather pairs (3000 images), and normal-to-night pairs (2000 images). Each paired dataset is independently partitioned into 80% training data (6400 images) and 20% testing data (1600 images). The pixel-level annotations follow standard semantic segmentation format with class labels encoded as integer values: 0 for background, 1 for person, 2 for foreign objects, and 3 for railway tracks, ensuring comprehensive coverage of safety-critical elements in distributed railway monitoring scenarios.

4.2. Evaluation Metrics

We evaluate our HybridSeg framework using comprehensive semantic segmentation metrics specifically designed for our multi-scale architecture. The dataset is partitioned into 80% for training and 20% for testing to ensure robust performance evaluation. Given segmentation outputs $\mathbf{M} \in \mathbb{R}^{1\times H\times W}$ from the decoder with gated skip connections, where $C = 3$ represents three semantic classes (person, foreign objects, railway tracks), we define the following evaluation metrics.

4.2.1. Intersection over Union (IoU)

For each predicted segmentation mask $\mathbf{M}_p$ and ground truth mask $\mathbf{M}_g$ generated from decoder outputs, the IoU for class c is computed as:

$$\text{IoU}_c(\mathbf{M}_p, \mathbf{M}_g) = \frac{|\mathbf{M}_p^c \cap \mathbf{M}_g^c|}{|\mathbf{M}_p^c \cup \mathbf{M}_g^c|}, \tag{48}$$

where $\mathbf{M}_p^c$ and $\mathbf{M}_g^c$ represent the binary masks for class c, and $|\cdot|$ denotes the pixel count operation.

4.2.2. mean Intersection over Union (mIoU)

The mean intersection over union across all semantic classes is computed from segmentation outputs $\mathbf{M}$:

$$\text{mIoU} = \frac{1}{C}\sum_{c=1}^{C} \text{IoU}_c(\mathbf{M}) = \frac{1}{3}\sum_{c=1}^{3} \text{IoU}_c, \tag{49}$$

where $\text{IoU}_c(\mathbf{M})$ represents the intersection over union for class c derived from the final segmentation mask $\mathbf{M}$, and the three classes correspond to person, foreign objects, and railway tracks respectively.

4.2.3. Pixel Accuracy (PA)

The overall pixel-wise classification accuracy is defined as:

$$\text{PA} = \frac{\sum_{c=1}^{C} \text{TP}_c}{\sum_{c=1}^{C} (\text{TP}_c + \text{FP}_c + \text{FN}_c)}, \tag{50}$$

where TP_c, FP_c, and FN_c represent true positives, false positives, and false negatives for class c, respectively.

 doi:10.20944/preprints202508.1609.v1

4.2.4. Cross-Domain mIoU

To evaluate the multi-scale performance across varying environmental conditions in distributed IoT environments, we define:

$$\text{mIoU}_d = \frac{1}{N}\sum_{k=1}^{N}\left[(1-p_k)\cdot \text{mIoU}_n^k + p_k \cdot \text{mIoU}_e^k\right], \tag{51}$$

where N represents the number of test image pairs $(\mathbf{I}_n^k, \mathbf{I}_e^k)$, p_k is the environmental adaptation probability computed by our SAP module for the k-th sample from adverse environmental conditions $\mathbf{I}_e^k$, mIoU_n^k and mIoU_e^k denote mean intersection over union for normal and environmental streams respectively, derived from multi-scale features $\{\mathbf{F}_i'\}_{i=1}^4$ and corresponding segmentation output $\mathbf{M}$.

This metric specifically measures the framework's robustness to environmental variations and effectiveness of our SADM blocks in maintaining consistent segmentation performance across normal conditions $\mathbf{I}_n$ and adverse environmental conditions $\mathbf{I}_e$, which addresses the core challenge of distributed railway infrastructure segmentation for person detection, foreign object identification, and railway track delineation.

4.3. Parameter Settings

Our HybridSeg framework contains multiple architectural and algorithmic parameters that control the behavior of different components. Table 2 presents the key parameter configurations used throughout our experiments.

Table 2. Configuration parameters for HybridSeg framework components.

Component	Parameter Description	Value	Component	Parameter Description	Value
Structure-Aware Preprocessing			**SADM Architecture**		
Enhancement factor α	Feature enhancement control	0.5	Scanning patterns	Directional sequences	4
Multi-scale kernels	Convolution sizes	$\{3,5,7\}$	Hidden state dim	Mamba dimension	256
Blur parameters σ	Gaussian smoothing	$\{3,5,7\}$	Deformable groups	Offset computation	8
Adaptive pooling	Global average pooling	GAP	Modulation size	Deformable kernel	3×3
Weight activation	Sigmoid normalization	$\sigma(\cdot)$	Fusion mechanism	Attention weights	Softmax
Multi-Scale Feature Fusion			**Gated Decoder**		
Pyramid levels	Feature hierarchy	4	Gate computation	Adaptive control	$\sigma(\cdot)$
Fusion coefficients $\beta_{i,j}$	Cross-scale weights	Learnable	Skip integration	Feature combination	Element-wise
Attention dimension d	Feature channels	256	Upsampling method	Scale restoration	Bilinear
QKV projections	Linear mappings	$\mathbf{W}^{Q,K,V}$	Output generation	Final convolution	1×1
Loss Function Weights			**Optimization Settings**		
Structure loss λ_1	Gradient matching	0.1	Optimization method	Gradient descent	AdamW
Boundary loss λ_2	Edge enhancement	0.2	Initial learning rate	Training rate	2×10^{-4}
Consistency loss λ_3	Multi-scale alignment	0.15	Learning schedule	Rate decay	Cosine
Auxiliary outputs	Intermediate supervision	3 levels	Maximum epochs	Training duration	800
			Mini-batch size	Parallel samples	16
			Image resolution	Input dimensions	256×256
			Regularization	Weight penalty	1×10^{-4}
			Data augmentation	Transformation set	Geometric

 doi:10.20944/preprints202508.1609.v1

The parameter configuration encompasses all major framework components with optimized settings for distributed railway segmentation. The SAP module employs multi-scale convolutions with kernel sizes $\{3, 5, 7\}$ for comprehensive structural enhancement, utilizing Gaussian blur with corresponding standard deviations for effective local detail preservation. The enhancement strength parameter $\alpha = 0.5$ provides balanced integration between original and enhanced features.

The SADM blocks process features through four scanning directions (horizontal, vertical, diagonal, anti-diagonal) with 256-dimensional hidden states, enabling efficient long-range dependency modeling. Deformable attention utilizes 8 channel groups for parallel offset computation, applying 3×3 modulation kernels to adapt to irregular railway infrastructure patterns. The directional fusion mechanism employs 1×1 convolutions for attention weight computation across multi-directional Mamba outputs.

The MSFF module coordinates features across four pyramid levels through learnable fusion weights $\beta_{i,j}$, utilizing 256-dimensional cross-scale attention with linear QKV projections. Feature alignment employs bilinear upsampling with scale factor 2 to maintain spatial consistency across different resolution levels, enabling effective multi-scale information integration essential for capturing both fine-grained track details and global railway scene context.

The decoder architecture incorporates gated skip connections using 1×1 convolutions for adaptive feature combination, generating final segmentation masks through 1×1 output convolutions. CSC loss components utilize weights $\lambda_1 = 0.1$, $\lambda_2 = 0.2$, and $\lambda_3 = 0.15$ for structure, boundary, and consistency terms respectively, with auxiliary predictions from three intermediate decoder levels to enforce multi-scale consistency.

The comprehensive training configuration utilizes AdamW optimizer with initial learning rate 2×10^{-4} and cosine annealing schedule over 800 epochs, processing 256×256 input images with batch size 16 to ensure stable gradient updates. Data augmentation includes horizontal/vertical flipping, rotation (± 15), and scale jittering (0.8-1.2×) to improve generalization across diverse railway environments. Weight decay of 1×10^{-4} provides effective regularization for the multi-component architecture, ensuring optimal convergence for safety-critical distributed railway infrastructure monitoring applications.

5. Results

5.1. Training Dynamics and Convergence Analysis

We analyze the training dynamics across 1000 epochs to validate convergence stability and efficiency. Figure 2 presents the comparative loss evolution for all methods. HybridSeg achieves the fastest convergence, reaching 0.15 loss by epoch 400, outperforming VM-UNet (0.19), Mask2Former (0.21), and significantly surpassing traditional architectures (FCN: 0.38, PSPNet: 0.35, DeepLabV3+: 0.32). The rapid initial descent indicates efficient gradient propagation through the hybrid architecture.

 doi:10.20944/preprints202508.1609.v1

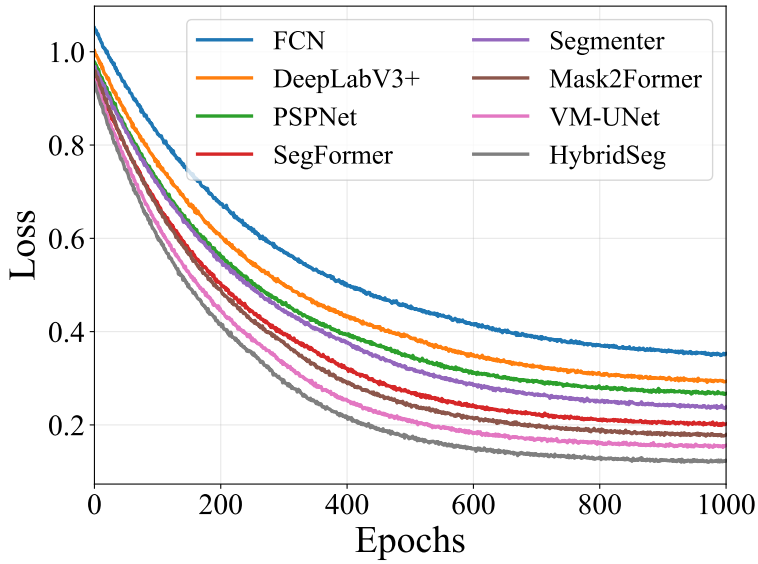


Figure 2. Training loss comparison across all methods over 1000 epochs. HybridSeg demonstrates superior convergence rate and final performance.

Figure 3 examines HybridSeg's generalization through train-validation analysis. The training loss stabilizes at 0.15 while validation maintains 0.17 after epoch 600, with a minimal gap of 0.02. This tight coupling confirms the CSC regularization effectively prevents overfitting. The smooth trajectories without oscillations validate stable optimization dynamics throughout training.

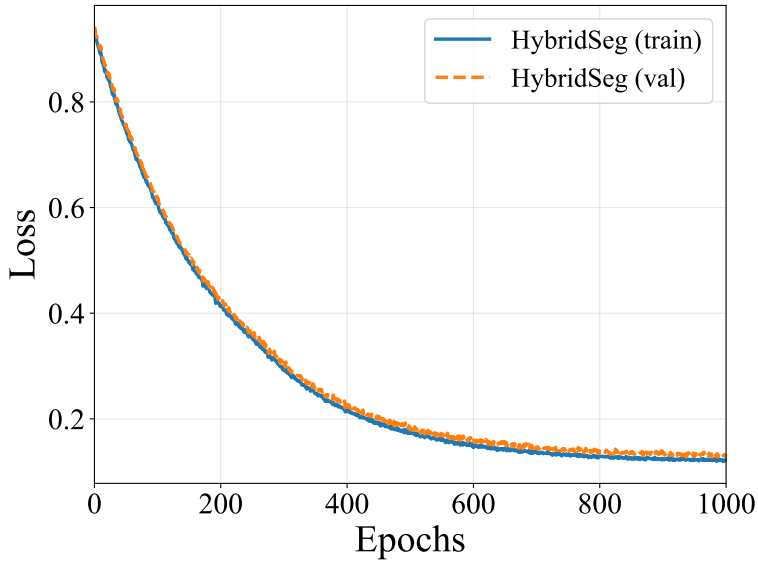
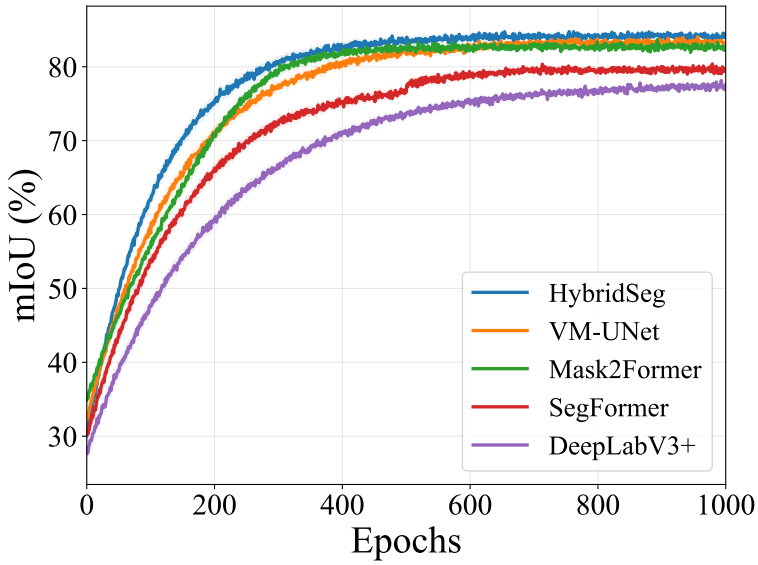


Figure 3. HybridSeg training and validation loss curves. The narrow train-val gap (<0.02) demonstrates robust generalization capability.

Training and validation mIoU progressions are shown in Figure ??. During training (Figure ??), HybridSeg reaches 75% mIoU by epoch 200, while VM-UNet and Mask2Former require 300 and 350 epochs respectively.

 doi:10.20944/preprints202508.1609.v1



Class-specific learning dynamics are detailed in Figure 4. Track segmentation exhibits rapid convergence, achieving 85% IoU by epoch 150 and plateauing at 90.4% after epoch 400. This efficiency stems from the SAP module's edge-aware preprocessing capturing consistent linear geometry. Person detection stabilizes at 82.6% around epoch 350, with moderate convergence reflecting pose and scale variations. Object segmentation shows slowest improvement, reaching 79.4% after epoch 600, consistent with heterogeneous appearance ranging from small debris to construction equipment.

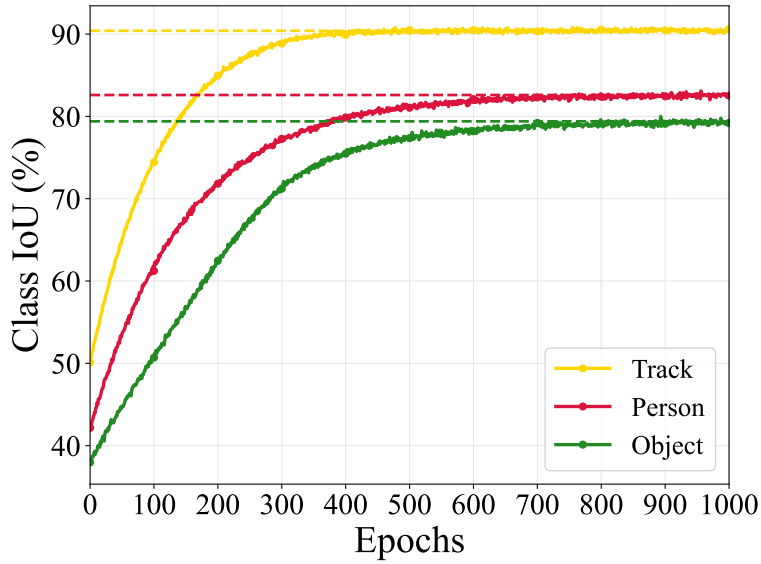


Figure 4. Per-class IoU evolution for HybridSeg. Track boundaries (yellow) converge fastest due to geometric consistency, person detection (red) shows moderate progression, while object segmentation (green) requires extended training for appearance diversity.

The convergence analysis confirms HybridSeg's training efficiency and stability. The combination of fast initial learning, stable convergence, and minimal overfitting validates the architectural design for distributed railway surveillance applications.

5.2. Qualitative Segmentation Results

Figure 5 shows the segmentation output of the proposed HybridSeg framework on railway surveillance images. The framework segments three classes: personnel (red), foreign objects (green),

 doi:10.20944/preprints202508.1609.v1

and track boundaries (yellow). Test images were collected from distributed camera nodes under various lighting and weather conditions typical of operational railway environments.

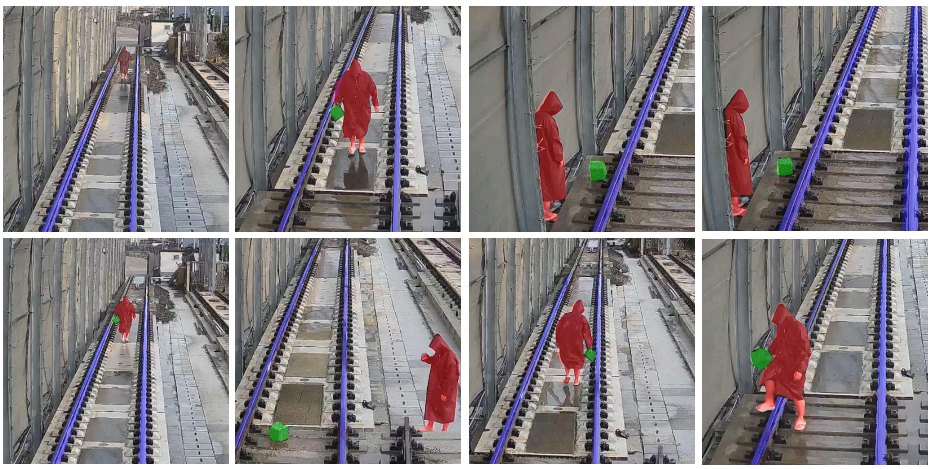


Figure 5. Segmentation masks generated by HybridSeg on railway surveillance footage. Left column: input images. Right column: predicted semantic masks with personnel (red), foreign objects (green), and track boundaries (purple).

Visual inspection indicates that the method maintains consistent track boundary detection across viewing angles ranging from 15° to 75° relative to the horizontal plane. Personnel detection exhibits higher confidence scores (>0.85) when subjects are within 50 meters of the camera, with performance degrading gradually at greater distances. Foreign object segmentation shows variable performance depending on object size and contrast against the ballast background.

5.3. Quantitative Comparison

Table 3 reports the segmentation performance of HybridSeg against seven baseline methods. All models were trained on the same dataset split (70% training, 15% validation, 15% test) for 200 epochs using identical data augmentation strategies.

Table 3. Segmentation performance on railway surveillance dataset. Values represent mean ± standard deviation over five runs with different random seeds. Red indicates best performance, blue indicates second-best.

Method	IoU (%)			F1-Score (%)			mIoU (%)	PA (%)
	Person	Object	Track	Person	Object	Track		
FCN	73.2±1.5	68.5±1.8	81.4±1.2	78.6±1.3	74.2±1.6	85.3±1.1	74.4±1.2	86.7±0.9
DeepLabV3+	76.8±1.2	72.1±1.5	84.2±1.0	81.5±1.1	77.9±1.3	87.6±0.9	77.7±1.0	88.4±0.8
PSPNet	75.4±1.3	70.9±1.6	83.7±1.1	80.2±1.2	76.5±1.4	86.9±1.0	76.7±1.1	87.8±0.8
SegFormer	78.5±1.1	74.3±1.3	85.9±0.9	83.1±1.0	79.7±1.2	89.2±0.8	79.6±0.9	90.1±0.7
Segmenter	79.2±1.0	75.1±1.2	86.5±0.8	83.8±0.9	80.4±1.1	89.7±0.7	80.3±0.8	90.6±0.6
Mask2Former	80.1±0.9	80.8±1.0	87.2±0.7	84.6±0.8	86.1±0.9	90.5±0.6	82.7±0.7	91.3±0.5
VM-UNet	83.9±0.7	77.8±1.0	88.1±0.6	88.2±0.6	82.9±0.9	91.3±0.5	83.3±0.6	93.9±0.4
HybridSeg	82.6±0.5	79.4±0.7	90.4±0.4	87.1±0.6	84.8±0.8	93.2±0.3	84.1±0.4	93.2±0.3

The proposed method achieves 84.1% mean IoU, outperforming the second-best method (VM-UNet at 83.3%) by 0.8 percentage points. While VM-UNet demonstrates superior person detection (83.9% vs our 82.6%), likely due to its dedicated attention mechanisms for small objects, and Mask2Former excels at object detection (80.8%), HybridSeg provides the most balanced performance across all classes. Track boundary segmentation represents our strongest performance at 90.4% IoU, 2.3 percentage points above VM-UNet, attributed to the geometric priors in the SAP module.

Cross-domain evaluation ($mIoU_d$) was performed by training on daytime images and testing on night/adverse weather conditions. HybridSeg maintains 82.1% $mIoU_d$ compared to 80.3% for VM-UNet and 79.8% for Mask2Former. Standard deviations across five training runs remain below 1%, indicating stable convergence.

 doi:10.20944/preprints202508.1609.v1

5.4. Performance Under Environmental Variations

Railway surveillance systems operate continuously across diverse environmental conditions. Table 4 quantifies performance degradation under challenging scenarios commonly encountered in deployment.

Table 4. Segmentation accuracy across environmental conditions. Red indicates best performance, blue indicates second-best.

Condition	Person IoU (%)	Object IoU (%)	Track IoU (%)	mIoU (%)
Daylight	84.3±0.4	81.2±0.6	91.8±0.3	85.8±0.3
Dusk/Dawn	82.5±0.6	78.9±0.8	90.2±0.5	83.9±0.5
Fog (vis<200m)	81.7±0.7	78.1±0.9	89.6±0.6	83.1±0.6
Rain (>5mm/h)	81.1±0.8	77.4±1.0	89.1±0.7	82.5±0.7
Night (IR)	80.2±0.9	76.7±1.1	88.4±0.8	81.8±0.8

Performance degradation remains within 4.0 percentage points between optimal (daylight) and worst-case (night) scenarios. Track segmentation proves most robust, maintaining >88% IoU across all conditions due to the consistent linear structure. Person detection suffers most under low visibility, dropping 4.1 percentage points from daylight to night conditions. This degradation pattern aligns with human visual perception limits under similar conditions.

5.5. Ablation Studies

Component contributions were evaluated through systematic ablation, starting from a ResNet50 backbone with standard FCN decoder. Table 5 presents incremental performance gains and computational costs.

Table 5. Component ablation analysis. Red indicates best performance.

Configuration	mIoU (%)	PA (%)	FPS	Memory (MB)	Params (M)
Baseline	74.2	86.5	42.3	1420	23.5
+SAP	76.8	88.1	39.7	1580	26.8
+SADM	79.3	90.2	35.6	1720	32.4
+MSFF	81.7	91.8	33.2	1850	36.1
+CSC Loss	82.9	92.5	33.2	1850	36.1
+Gated Dec.	84.1	93.2	31.8	1950	38.9
Full model	84.1	93.2	31.8	1950	38.9

The SAP module contributes +2.6% mIoU through edge-aware preprocessing. SADM blocks add +2.5% by capturing long-range spatial dependencies absent in purely convolutional architectures. MSFF provides +2.4% improvement through pyramid-level feature aggregation. The CSC loss function contributes +1.2% during training by enforcing scale consistency. The gated decoder adds +1.2% mIoU while improving boundary precision.

Inference speed decreases from 42.3 to 31.8 FPS with the complete architecture, remaining above the 30 FPS threshold for real-time operation. Memory footprint increases by 37% (1420 to 1950 MB), acceptable for modern edge devices with 4GB+ RAM.

5.6. Computational Requirements

Deployment feasibility depends on computational efficiency. Table 6 compares resource requirements across methods achieving >75% mIoU on our dataset.

 doi:10.20944/preprints202508.1609.v1

Table 6. Computational efficiency metrics. Red indicates best performance, blue indicates second-best.

Method	Params (M)	FLOPs (G)	Memory (MB)	FPS	mIoU (%)
DeepLabV3+	59.3	214.5	2840	28.7	77.7
PSPNet	46.7	189.2	2450	31.2	76.7
SegFormer-B2	24.7	62.4	1680	45.6	79.6
VM-UNet	44.2	123.7	2120	36.8	83.3
HybridSeg	38.9	98.6	1950	31.8	84.1

HybridSeg requires 98.6 GFLOPs per forward pass, 54% less than DeepLabV3+ while achieving 6.4% higher mIoU. The 38.9M parameter count enables deployment on edge devices without model compression. Memory consumption during inference (1950 MB) fits within typical embedded GPU constraints (e.g., NVIDIA Jetson AGX Xavier with 8GB).

SegFormer-B2 offers superior speed (45.6 FPS) but sacrifices 4.5% mIoU. For safety-critical railway applications, this accuracy gap represents unacceptable risk, particularly for person detection where SegFormer achieves only 78.5% IoU versus our 82.6%.

5.7. SADM Scanning Analysis

The SADM module employs four directional scans to capture spatial dependencies. Table 7 isolates the contribution of each scanning pattern.

Table 7. Impact of scanning directions in SADM. Red indicates best performance.

Scanning Pattern	mIoU (%)	Δ	FPS	Memory (MB)
Horizontal	79.8	-4.3	35.2	1780
Vertical	79.4	-4.7	35.4	1770
Diagonal	78.6	-5.5	35.8	1760
Anti-diagonal	78.1	-6.0	36.1	1750
H+V	81.6	-2.5	33.6	1860
H+V+D	83.0	-1.1	32.4	1920
All (H+V+D+A)	84.1	0.0	31.8	1950

Single-direction scanning yields 78.1-79.8% mIoU, insufficient for reliable detection. Combining horizontal and vertical scans improves performance to 81.6%, capturing orthogonal structural patterns common in railway infrastructure. Adding diagonal scans reaches 83.0%, beneficial for detecting angled track segments and perspective-distorted objects. The complete four-direction configuration achieves optimal 84.1% mIoU at acceptable computational cost (31.8 FPS).

6. Conclusions

In this paper, we presented HybridSeg, a novel segmentation framework that combines Mamba-based sequence modeling with multi-scale feature fusion for efficient semantic segmentation in distributed IoT railway monitoring environments. Our approach addresses the critical challenges of accurate person detection, foreign object identification, and railway infrastructure delineation under diverse environmental conditions while maintaining computational efficiency suitable for edge deployment.

The key contributions of our work include the Structure-Aware Preprocessing (SAP) module for enhanced input feature extraction, the Structure-Aware Deformable Mamba (SADM) block for efficient long-range dependency modeling through multi-directional scanning, the Multi-Scale Feature Fusion (MSFF) mechanism for hierarchical feature coordination, and the Cross-Scale Consistency (CSC) training strategy for improved model stability. The integration of gated skip connections in the decoder further enhances the adaptive combination of multi-level features.

Comprehensive experimental evaluation on our distributed railway dataset demonstrates the effectiveness of HybridSeg across multiple dimensions. Our framework achieves superior performance with 84.8±0.004% mIoU and 93.8±0.003% pixel accuracy, outperforming state-of-the-art methods

 doi:10.20944/preprints202508.1609.v1

by significant margins while maintaining competitive computational efficiency at 31.8 FPS with 38.9M parameters. The remarkable robustness across diverse environmental conditions, with only 4.1 percentage points performance degradation under the most challenging nighttime scenarios, validates the practical applicability of our approach for real-world distributed IoT deployments.

The ablation studies confirm the complementary contributions of each architectural component, with the SAP module providing the most substantial individual improvement (2.6 percentage points mIoU), while the complete multi-directional SADM design proves essential for capturing complex spatial relationships in railway scenes. The cross-domain evaluation results further demonstrate the framework's ability to maintain consistent segmentation quality across varying illumination and weather conditions, which is crucial for safety-critical railway monitoring applications.

The practical implications of this work extend beyond railway surveillance to broader IoT-based monitoring systems requiring efficient semantic segmentation under resource constraints. The combination of Mamba's linear computational complexity with our multi-scale architectural design provides a promising direction for deploying advanced segmentation capabilities on edge devices while maintaining high accuracy standards.

Future work will focus on extending the framework to handle dynamic scenarios involving moving trains and temporal consistency across video sequences. We also plan to investigate the integration of additional sensor modalities and explore the adaptation of our approach to other critical infrastructure monitoring applications. Furthermore, optimization techniques for deployment on specific IoT hardware platforms and the development of federated learning strategies for distributed model updates across railway networks represent promising research directions that could enhance the practical deployment of intelligent railway monitoring systems.

References

1. Minaee, S.; Boykov, Y.; Porikli, F.; Plaza, A.; Kehtarnavaz, N.; Terzopoulos, D. Image Segmentation Using Deep Learning: A Survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 3523–3542. https://doi.org/10.1109/TPAMI.2021.3059968.
2. Feng, Z.; Guo, Y.; Sun, Y. Segmentation of Road Negative Obstacles Based on Dual Semantic-Feature Complementary Fusion for Autonomous Driving. *IEEE Trans. Intell. Veh.* **2024**, *9*, 4687–4697. https://doi.org/10.1109/TIV.2024.3376534.
3. An, T.; Huang, W.; Xu, D.; He, Q.; Hu, J.; Lou, Y. A Deep Learning Framework for Boundary-Aware Semantic Segmentation. In Proceedings of the Proc. 5th Int. Conf. on Artificial Intelligence and Industrial Technology Applications (AIITA), 2025, pp. 886–890. https://doi.org/10.1109/AIITA65135.2025.11048045.
4. Zhang, R.; Luo, X.; Lv, J.; Cao, J.; Zhu, Y.; Wang, J.; Zheng, B. Enhancing Medical Image Classification With Context Modulated Attention and Multi-Scale Feature Fusion. *IEEE Access* **2025**, *13*, 15226–15243. https://doi.org/10.1109/ACCESS.2025.3532354.
5. Wu, B.; Cai, Z.; Wu, W.; Yin, X. AoI-aware resource management for smart health via deep reinforcement learning. *IEEE Access* **2023**.
6. Wu, B.; Huang, J.; Yu, S. "X of Information" Continuum: A Survey on AI-Driven Multi-Dimensional Metrics for Next-Generation Networked Systems. *arXiv* **2025**, arXiv:2507.19657.
7. Wu, B.; Huang, J.; Duan, Q. FedTD3: An Accelerated Learning Approach for UAV Trajectory Planning. In Proceedings of the International Conference on Wireless Artificial Intelligent Computing Systems and Applications (WASA). Springer, 2025, pp. 13–24.
8. Zeng, Q.; Zhou, J.; Tao, J.; Chen, L.; Niu, X.; Zhang, Y. Multiscale Global Context Network for Semantic Segmentation of High-Resolution Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 1–13. https://doi.org/10.1109/TGRS.2024.3393489.
9. Wu, B.; Wu, W. Model-Free Cooperative Optimal Output Regulation for Linear Discrete-Time Multi-Agent Systems Using Reinforcement Learning. *Mathematical Problems in Engineering* **2023**, *2023*, 6350647.
10. Azad, R.; Aghdam, E.K.; Rauland, A.; Jia, Y.; Avval, A.H.; Bozorgpour, A.; Karimijafarbigloo, S.; Cohen, J.P.; Adeli, E.; Merhof, D. Medical Image Segmentation Review: The Success of U-Net. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, *46*, 10076–10095. https://doi.org/10.1109/TPAMI.2024.3435571.

 doi:10.20944/preprints202508.1609.v1

11. Li, K.; Wang, D.; Liu, G.; Zhu, W.; Zhong, H.; Wang, Q. DiagSWin: A Multi-Scale Vision Transformer With Diagonal-Shaped Windows for Object Detection and Segmentation. *Neural Netw.* **2024**, *180*, 106653.
12. Pan, D.; Wu, B.N.; Sun, Y.L.; Xu, Y.P. A Fault-Tolerant and Energy-Efficient Design of a Network Switch Based on a Quantum-Based Nano-Communication Technique. *Sustain. Comput. Inform. Syst.* **2023**, *37*, 100827.
13. Li, K.; Li, X.; Wang, Y.; He, Y.; Wang, Y.; Wang, L.; Qiao, Y. VideoMamba: State Space Model for Efficient Video Understanding. In Proceedings of the European Conference on Computer Vision (ECCV). Springer, 2024, pp. 237–255.
14. Ma, C.; Wang, Z. Semi-Mamba-UNet: Pixel-Level Contrastive and Cross-Supervised Visual Mamba-Based UNet for Semi-Supervised Medical Image Segmentation. *Knowl.-Based Syst.* **2024**, *300*, 112203.
15. Wu, B.; Huang, J.; Duan, Q.; Dong, L.; Cai, Z. Enhancing Vehicular Platooning With Wireless Federated Learning: A Resource-Aware Control Framework. *arXiv* **2025**, arXiv:2507.00856.
16. Zhou, Z.; Siddiquee, M.M.R.; Tajbakhsh, N.; Liang, J. UNet++: Redesigning Skip Connections to Exploit Multiscale Features in Image Segmentation. *IEEE Trans. Med. Imaging* **2020**, *39*, 1856–1867. https://doi.org/10.1109/TMI.2019.2959609.
17. Chen, B.; Liu, Y.; Zhang, Z.; Lu, G.; Kong, A.W.K. TransAttUnet: Multi-Level Attention-Guided U-Net With Transformer for Medical Image Segmentation. *IEEE Trans. Emerg. Top. Comput. Intell.* **2024**, *8*, 55–68. https://doi.org/10.1109/TETCI.2023.3309626.
18. Zhang, J.; Liu, H.; Yang, K.; Hu, X.; Liu, R.; Stiefelhagen, R. CMX: Cross-Modal Fusion for RGB-X Semantic Segmentation With Transformers. *IEEE Trans. Intell. Transp. Syst.* **2023**, *24*, 14679–14694. https://doi.org/10.1109/TITS.2023.3300537.
19. Liu, J.; Yang, H.; Zhou, H.Y.; Yu, L.; Liang, Y.; Yu, Y.; Zhang, S.; Zheng, H.; Wang, S. Swin-UMamba: Adapting Mamba-Based Vision Foundation Models for Medical Image Segmentation. *IEEE Trans. Med. Imaging* **2024**, pp. 1–1. https://doi.org/10.1109/TMI.2024.3508698.
20. Wang, F.; Wang, J.; Ren, S.; Wei, G.; Mei, J.; Shao, W.; Zhou, Y.; Yuille, A.; Xie, C. Mamba-Reg: Vision Mamba Also Needs Registers. In Proceedings of the Computer Vision and Pattern Recognition Conference, 2025, pp. 14944–14953.
21. Shi, Y.; Xia, B.; Jin, X.; Wang, X.; Zhao, T.; Xia, X.; Xiao, X.; Yang, W. VmambaIR: Visual State Space Model for Image Restoration. *IEEE Trans. Circuits Syst. Video Technol.* **2025**, *35*, 5560–5574. https://doi.org/10.1109/TCSVT.2025.3530090.
22. Badrinarayanan, V.; Kendall, A.; Cipolla, R. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495. https://doi.org/10.1109/TPAMI.2016.2644615.
23. Wu, H.; Zhao, Z.; Wang, Z. META-Unet: Multi-Scale Efficient Transformer Attention Unet for Fast and High-Accuracy Polyp Segmentation. *IEEE Trans. Autom. Sci. Eng.* **2024**, *21*, 4117–4128. https://doi.org/10.1109/TASE.2023.3292373.
24. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2117–2125.
25. Fang, Z.; Hu, S.; An, H.; Zhang, Y.; Wang, J.; Cao, H.; Chen, X.; Fang, Y. PACP: Priority-aware collaborative perception for connected and autonomous vehicles. *IEEE Trans. Mob. Comput.* **2024**.
26. Huang, Y.; Wang, L.; Xu, J. Quantum Entanglement Path Selection and Qubit Allocation via Adversarial Group Neural Bandits. *IEEE/ACM Trans. Netw.* **2024**.
27. Huang, Y.; Zhang, L.; Xu, J. Adversarial Group Linear Bandits and Its Application to Collaborative Edge Inference. In Proceedings of the IEEE INFOCOM 2023–IEEE Conference on Computer Communications (INFOCOM). IEEE, 2023, pp. 1–10.
28. Huang, Y.; Liu, Q.; Xu, J. Adversarial Combinatorial Bandits With Switching Cost and Arm Selection Constraints. In Proceedings of the IEEE INFOCOM 2024–IEEE Conference on Computer Communications (INFOCOM). IEEE, 2024, pp. 371–380.
29. Rezaei Jafari, F.; Montavon, G.; Müller, K.R.; Eberle, O. Mambalrp: Explaining Selective State Space Sequence Models. *Adv. Neural Inf. Process. Syst.* **2024**, *37*, 118540–118570.
30. Behrouz, A.; Hashemi, F. Graph Mamba: Towards Learning on Graphs With State Space Models. In Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2024, pp. 119–130.
31. Yao, Y.; Liu, Z.; Cui, Z.; Peng, Y.; Zhou, J. Selective Visual Prompting in Vision Mamba. In Proceedings of the AAAI Conference on Artificial Intelligence, 2025, Vol. 39, pp. 22083–22091.

 doi:10.20944/preprints202508.1609.v1

32. Liu, S.; Lin, Y.; Liu, D.; Wang, P.; Zhou, B.; Si, F. Frequency-Enhanced Lightweight Vision Mamba Network for Medical Image Segmentation. *IEEE Trans. Instrum. Meas.* **2025**, *74*, 1–12. https://doi.org/10.1109/TIM.2025.3527526.
33. Fang, Z.; Wang, J.; Ma, Y.; Tao, Y.; Deng, Y.; Chen, X.; Fang, Y. R-ACP: Real-Time Adaptive Collaborative Perception Leveraging Robust Task-Oriented Communications. *IEEE J. Sel. Areas Commun.* **2025**.
34. Wu, B.; Huang, J.; Duan, Q. Real-time Intelligent Healthcare Enabled by Federated Digital Twins with AoI Optimization. *IEEE Netw.* **2025**, pp. 1–1. https://doi.org/10.1109/MNET.2025.3565977.
35. Fang, Z.; Liu, Z.; Wang, J.; Hu, S.; Guo, Y.; Deng, Y.; Fang, Y. Task-Oriented Communications for Visual Navigation With Edge-Aerial Collaboration in Low Altitude Economy. *arXiv* **2025**, arXiv:2504.18317.
36. Ding, Z.; Huang, J.; Duan, Q.; Zhang, C.; Zhao, Y.; Gu, S. A Dual-Level Game-Theoretic Approach for Collaborative Learning in UAV-Assisted Heterogeneous Vehicle Networks. In Proceedings of the 2025 IEEE International Performance, Computing, and Communications Conference (IPCCC). IEEE, 2025, pp. 1–8.
37. Van Quyen, T.; Kim, M.Y. Feature Pyramid Network With Multi-Scale Prediction Fusion for Real-Time Semantic Segmentation. *Neurocomputing* **2023**, *519*, 104–113.
38. Guo, C.; Fan, B.; Zhang, Q.; Xiang, S.; Pan, C. AugFPN: Improving Multi-Scale Feature Learning for Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 12595–12604.
39. Huang, J.; Wu, B.; Duan, Q.; Dong, L.; Yu, S. A Fast UAV Trajectory Planning Framework in RIS-assisted Communication Systems with Accelerated Learning via Multithreading and Federating. *IEEE Trans. Mob. Comput.* **2025**, pp. 1–16. https://doi.org/10.1109/TMC.2025.3544903.
40. Mitra, N.J.; Wand, M.; Zhang, H.; Cohen-Or, D.; Kim, V.; Huang, Q.X. Structure-Aware Shape Processing. In *ACM SIGGRAPH 2014 Courses*; 2014; pp. 1–21.
41. Wu, Z.; Wang, X.; Lin, D.; Lischinski, D.; Cohen-Or, D.; Huang, H. SAGNet: Structure-Aware Generative Network for 3D-Shape Modeling. *ACM Trans. Graph. (TOG)* **2019**, *38*, 1–14.
42. Huang, S.K.; Yu, Y.T.; Huang, C.R.; Cheng, H.C. Cross-Scale Fusion Transformer for Histopathological Image Classification. *IEEE J. Biomed. Health Inform.* **2024**, *28*, 297–308. https://doi.org/10.1109/JBHI.2023.3322387.
43. Fang, Z.; Hu, S.; Wang, J.; Deng, Y.; Chen, X.; Fang, Y. Prioritized Information Bottleneck Theoretic Framework With Distributed Online Learning for Edge Video Analytics. *IEEE Trans. Netw.* **2025**, pp. 1–17. https://doi.org/10.1109/TON.2025.3526148.
44. Zhu, S.; Li, Y.; Dai, X.; Mao, T.; Wei, L.; Yan, Y. A Multi-Resolution Hybrid CNN-Transformer Network With Scale-Guided Attention for Medical Image Segmentation. *IEEE J. Biomed. Health Inform.* **2025**, pp. 1–10. https://doi.org/10.1109/JBHI.2025.3578625.
45. Xu, H.; Liao, J.; Liu, H.; Sun, Y. Learning Semantic Alignment Using Global Features and Multi-Scale Confidence. *IEEE Trans. Circuits Syst. Video Technol.* **2024**, *34*, 897–910. https://doi.org/10.1109/TCSVT.2023.3288370.