Lightweight Network-Based Semantic Segmentation for UAVs and Its RISC-V Implementation

Yankai Chen * , Hongkun Du , Yutong Zhou

Posted Date: 14 August 2025

doi: 10.20944/preprints202508.1108.v1

Keywords: U-Net; MobileNetV2; semantic segmentation; RISC-V

 doi:10.20944/preprints202508.1108.v1

Article

Lightweight Network-Based Semantic Segmentation for UAVs and its RISC-V Implementation

Yankai Chen [1,*], Hongkun Du [2] and Yutong Zhou [3]

[1] School of Mathematics and Statistics, Ningbo University, Ningbo, China
[2] College of Science and Engineering, Flinders University, Adelaide, Australia
[3] School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, United States
* Correspondence: 2813745942@qq.com

Abstract

This study proposes a lightweight solution for real-time, low-power semantic segmentation in drone applications based on a RISC-V heterogeneous architecture. A lightweight model integrating U-Net and MobileNetV2 was designed to maintain multi-scale feature extraction capabilities while reducing computational complexity to one-tenth that of standard convolutions through depthwise separable convolutions. Leveraging the modular flexibility of the RISC-V processor and the hardware acceleration capabilities of FPGA, a heterogeneous computing framework was established, supporting customized instructions (e.g., DCONV, TCONV) and on-chip SRAM tiling optimization. Experimental results demonstrate that the model achieves an mIoU of approximately 0.2 and 70% pixel-level accuracy on the validation set, with inference latency reduced by 3–5 times via FPGA acceleration. The TrustZone module ensures secure model deployment through SM3/SM4 cryptographic validation. This work provides a scalable open-source framework for high-precision semantic segmentation on edge devices, validating the engineering feasibility of RISC-V in vision-centric edge intelligence.

Keywords: U-Net; MobileNetV2; semantic segmentation; RISC-V

1. Introduction

With the rapid development of drone technology, its applications in fields such as agricultural monitoring, urban planning, and disaster response have expanded significantly. High-resolution drone imagery offers abundant scene information [1], yet also imposes demanding requirements for multi-scale, multi-category object recognition, challenging traditional semantic segmentation algorithms [2]. Existing segmentation methods often rely on high-performance hardware (e.g., GPUs or specialized AI chips), limiting their suitability for resource-constrained edge devices on drones. This creates a critical need for lightweight, efficient models capable of real-time processing with low power consumption [3].

RISC-V, as a modular, open-source instruction set architecture (ISA), is gaining traction in the edge computing domain due to its scalability and energy efficiency. Its customizable instruction set makes it particularly suitable for embedded vision tasks, enabling tailored acceleration for specific operations. However, deploying complex vision models like semantic segmentation on RISC-V platforms remains a challenge due to constraints in computation, memory bandwidth, and co-design optimization between hardware and algorithms [4–6].

This study addresses the core requirements of real-time performance and low power consumption in semantic segmentation tasks for drone applications by proposing a lightweight solution based on a RISC-V heterogeneous architecture. The goal is to establish an efficient algorithm-hardware co-optimization framework for edge computing. First, in response to the challenges of multi-scale, multi-class object recognition in high-resolution drone imagery processing, a lightweight

 doi:10.20944/preprints202508.1108.v1

model integrating U-Net and MobileNetV2 is designed. By leveraging depthwise separable convolutions, computational complexity is significantly reduced while retaining multi-scale feature extraction capabilities, thereby enhancing model efficiency. Second, at the hardware level, a heterogeneous architecture combining a RISC-V core and an FPGA accelerator is constructed. This design incorporates customized instruction sets and on-chip SRAM tiling strategies to accelerate convolution operations. Additionally, image tiling reduces energy consumption from external memory access, effectively alleviating bandwidth bottlenecks. Furthermore, a TrustZone security module and SM3/SM4 cryptographic mechanisms are introduced to ensure the security and integrity of model deployment. Finally, experiments validate the model's performance, demonstrating reduced inference latency through FPGA acceleration. A systematic evaluation of the algorithm-hardware co-design confirms its feasibility and performance advantages on edge devices, providing theoretical support and engineering paradigms for the practical application of RISC-V in autonomous drone vision systems.

2. Related Work

2.1. Drone Semantic Segmentation

In recent years, scholars worldwide have conducted numerous studies in the field of drone semantic segmentation, proposing various improved models and methods for diverse application scenarios. Liu et al. introduced a dense small-object segmentation approach for drone imagery based on GSegFormer, which enhances segmentation accuracy for small targets through a multi-scale low-loss feature fusion network and a Cascaded Gated Attention Module (CGAM). This method proves particularly effective for processing densely distributed small objects in aerial imagery, though its complex structure may incur higher computational costs. Girisha et al. validated the feasibility of FCN and U-Net architectures for green vegetation and road segmentation using manually annotated drone video datasets, achieving pixel accuracies of 89.7% and 87.31% respectively7. However, their study was limited by a small dataset scale and did not address multi-category segmentation in complex backgrounds. Xiong's team developed the Tea-UNet model, integrating a RepViT encoder with Multi-level Feature Transform (MFT) and Multi-scale Attention Fusion (MAF) modules. This solution achieved high-precision segmentation in tea plantation drone imagery, especially excelling for elongated or irregularly shaped targets—though its optimization strategy requires further validation for generalizability to other agricultural scenarios. Focusing on disaster assessment, Pi et al. employed Mask-RCNN and PSPNet on the Volan2019 dataset, improving robustness in post-disaster damage detection through multi-class segmentation and targeted data augmentation. Nevertheless, the models showed reduced accuracy for small objects (e.g., vehicles) and relied on balanced datasets with high annotation costs. Collectively, current research predominantly addresses small-object segmentation, multi-scale feature fusion, and adaptability to complex scenes. Yet persistent limitations include restricted dataset scales, insufficient computation-accuracy tradeoffs, and scenario-specific dependencies. Future efforts should prioritize lightweight model design, multimodal data fusion, and adaptive augmentation strategies to enhance generalization capabilities.

2.2. RISC-V

Recent years have witnessed significant advancements in RISC-V-based embedded system design by scholars globally, yielding innovative solutions for diverse application scenarios. Liu et al. developed an intelligent cold-storage monitoring system using RISC-V SoC CH2601 with the Alibaba Cloud IoT platform, integrating multisensory and remote control modules to enable real-time monitoring and automated regulation of environmental parameters (temperature, humidity, UV intensity). Key advantages include AT command-based cloud platform interactions and smartphone remote control with strong extensibility, though its multitasking capability in complex scenarios is constrained by hardware resource allocation strategies8. Meanwhile, Wang et al. created a Chisel-based RISC-V five-stage pipelined microcontroller for controlled nuclear fusion applications.

 doi:10.20944/preprints202508.1108.v1

Through data forwarding and branch prediction optimizations, it achieves 2.55 CoreMarks/MHz while successfully running RT-Thread on FPGA9. However, full stability validation under high-radiation environments remains lacking, and compatibility with non-standard ISA extensions requires further evaluation. Collectively, contemporary research emphasizes RISC-V's customizable strengths and IoT integration capabilities, yet reveals improvement opportunities in heterogeneous resource scheduling, extreme-environment adaptability, and industrial-grade reliability verification. Future efforts should leverage advanced process nodes and fault-tolerant mechanisms to extend its application boundaries.

3. Research Design

3.1. U-Net + MobileNetV2 Architecture for Semantic Segmentation

The proposed semantic segmentation model adopts a U-Net architecture with MobileNetV2 as the encoder backbone, designed to balance accuracy and computational efficiency for drone imagery. The encoder-decoder framework employs multi-scale feature extraction and skip connections to preserve spatial resolution while capturing semantic context.

The encoder is constructed using MobileNetV2, which integrates linear bottleneck layers and inverted residual blocks. Each block consists of three stages:

Expansion Layer: Expands input channels via a 1×1 convolution. For an input tensor $X \in \mathbb{R}^{H\times W\times C_{in}}$, the expanded feature map X_{expand} is:

$$X_{expand} = \mathrm{ReLU6}\left(W_{expand} * X + b_{expand}\right) \tag{1}$$

where $W_{expand} \in \mathbb{R}^{1\times 1\times C_{in}\times C_{expand}}$ is the weight matrix and b_{expand} is the bias.

Depthwise Convolution: Applies spatial filtering independently to each channel:

$$X_{depth} = \mathrm{ReLU6}\left(W_{depth} * X_{expand} + b_{depth}\right) \tag{2}$$

where $W_{depth} \in \mathbb{R}^{k\times k\times C_{expand}\times 1}$ is the depthwise kernel.

Projection Layer: Reduces channel dimensions via another 1×1 convolution:

$$X_{project} = \mathrm{ReLU6}\left(W_{project} * X_{depth} + b_{project}\right) \tag{3}$$

where $W_{project} \in \mathbb{R}^{1\times 1\times C_{expand}\times C_{out}}$.

This design reduces computational cost while maintaining feature richness through depthwise separable convolutions.

The decoder progressively upscales features to recover spatial resolution. For the l-th layer:

Upsampling: Uses transposed convolution to double the feature map size:

$$U_l = \mathrm{UpSample}(D_l) \in \mathbb{R}^{2H_l\times 2W_l\times C_l} \tag{4}$$

where UpSample applies a 2×2 kernel with stride=2.

Feature Fusion: Concatenates upsampled features U_l with corresponding encoder outputs F_l along the channel dimension:

$$D_{l-1} = \mathrm{Concat}(U_l, F_l) \in \mathbb{R}^{2H_l\times 2W_l\times 2C_l} \tag{5}$$

Refinement: Applies 3×3 convolution to refine fused features:

$$D_{l-1} = \mathrm{Conv}_{3\times 3}(D_{l-1}) \in \mathbb{R}^{2H_l\times 2W_l\times C_{out}} \tag{6}$$

The decoder outputs are passed through a 1×1 convolution to map features to $N_{class} = 23$ semantic classes:

$$\hat{Y} = \mathrm{Conv}_{1\times 1}(D_0) \in \mathbb{R}^{H_{input}\times W_{input}\times N_{class}} \tag{7}$$

Loss Function: Cross-entropy loss minimizes the discrepancy between predicted $\hat{Y}$ and ground truth Y:

 doi:10.20944/preprints202508.1108.v1

$$\mathcal{L}_{CE} = -\frac{1}{N}\sum_{i=1}^{H}\sum_{j=1}^{W}\sum_{c=1}^{N_{class}} Y_{ijc}\log\left(\hat{Y}_{ijc}\right) \tag{8}$$

where $N = H \times W \times N_{class}$.

Optimizer: AdamW with OneCycleLR adjusts learning rates dynamically. Parameter updates follow:

$$\theta_{t+1} = \theta_t - \eta_t \cdot \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon} - \eta_t \cdot \lambda \cdot \theta_t \tag{9}$$

where $\hat{m}_t, \hat{v}_t$ are bias-corrected momentum and variance terms, η_t is the learning rate, and λ is weight decay.

Learning Rate Schedule: OneCycleLR cyclically varies η_t between η_{min} and η_{max}:

$$\eta_t = \begin{cases} \eta_{min} + (\eta_{max} - \eta_{min}) \cdot \dfrac{t}{T_{cycle}} & \text{(ascending phase)} \\ \eta_{max} - (\eta_{max} - \eta_{min}) \cdot \dfrac{t - \frac{T_{cycle}}{2}}{\frac{T_{cycle}}{2}} & \text{(descending phase)} \end{cases} \tag{10}$$

Depthwise Separable Convolution Efficiency: Reduces parameters and FLOPs compared to standard convolutions:

Standard Convolution:

$$\text{Params}_{std} = k^2 \cdot C_{in} \cdot C_{out}, \text{FLOPs}_{std} = 2 \cdot k^2 \cdot C_{in} \cdot C_{out} \cdot H \cdot W \tag{11}$$

Tiling for Memory Efficiency: High-resolution images (4000×6000) are divided into 512×768 tiles to fit on-chip memory:

$$X_i = X_{(i \cdot h:(i+1) \cdot h, j \cdot w:(j+1) \cdot w, :)} \text{ for } i, j \tag{12}$$

3.2. Hardware Architecture Design for RISC-V Heterogeneous Processors

To achieve real-time inference for semantic segmentation on a RISC-V heterogeneous processor, the architecture integrates a RISC-V core with FPGA-based acceleration units. The RISC-V core manages control flow and logical operations, while the FPGA handles computationally intensive tasks such as convolution and activation functions.

The RISC-V core communicates with the FPGA accelerator via an AXI bus. The FPGA embeds a customized convolution accelerator (Convolution Accelerator) optimized for depthwise separable convolutions. The data flow is defined as:

$$X_{out} = \text{FPGA\_Accelerator}\left(X_{in}, W_{dw}, W_{proj}\right) \tag{13}$$

where W_{dw} and W_{proj} denote the weight matrices of depthwise and pointwise convolutions, respectively. The FPGA employs a pipelined architecture to enable parallel processing across multiple channels. Each channel's computation is formulated as:

$$Y_i^{(t)} = \sum_{k=1}^{K_h}\sum_{l=1}^{K_w} X_i^{(t-k)} \cdot W_{dw}(k, l), i = 1, 2, \ldots, C_{in} \tag{14}$$

Here, t represents the time step, and $Y_i^{(t)}$ is the output feature of the i_{th} channel.

To reduce external memory access latency, on-chip SRAM caches weights and intermediate feature maps. High-resolution images (e.g., 1024×1024) are partitioned into 64×64 tiles, mapped to on-chip memory as:

$$\text{Tile}_{m,n} = X[m \cdot 64: (m+1) \cdot 64, n \cdot 64: (n+1) \cdot 64, :], m, n \in \mathbb{Z}^+ \tag{15}$$

 doi:10.20944/preprints202508.1108.v1

Each tile is processed independently in the FPGA, minimizing cross-bank memory access energy.

The RISC-V core integrates a TrustZone module to ensure model integrity. During boot-up, the SM3 hash algorithm measures the bootloader, OS, and model weights:

$$\text{Digest} = \text{SM3}(\text{Data}), \text{Signature} = \text{SM4}(\text{Digest},\text{Key}) \tag{16}$$

If the computed digest matches the pre-stored signature, the model is loaded; otherwise, a security exception is triggered.

3.3. Customized Instruction Set Extensions and Hardware Acceleration

RISC-V's open architecture enables custom instruction extensions for critical operations. Key instructions for semantic segmentation include:

This instruction executes depthwise separable convolution directly on the FPGA. Its format is:

$$\text{DCONV rd,rs1,rs2,imm} \tag{17}$$

Here, rs1 holds the input feature address, rs2 the weight address, imm specifies kernel size and stride, and rd stores the output. The FPGA loads input features and weights into local memory, computing results via a Processing Element (PE) array.

Upsampling in the decoder is accelerated via transposed convolution, defined as:

$$Y_{i,j}^{(k)} = \sum_{m=1}^{K_h} \sum_{n=1}^{K_w} W_{m,n}^{(k)} \cdot X_{\left\lfloor \frac{i}{s} \right\rfloor - m+1, \left\lfloor \frac{j}{s} \right\rfloor - n+1} \tag{18}$$

The TCONV instruction leverages the FPGA's transposed convolution engine to reduce core computational load.

4. Experimental Result

Next, the results of drone (UAV) image semantic segmentation based on the U-net + MobileNetV2 architecture will be presented here. Specifically, the loss iteration over batches, the mean Intersection over Union (mIoU) score, and the accuracy curves are plotted as shown in Figures 1–3.

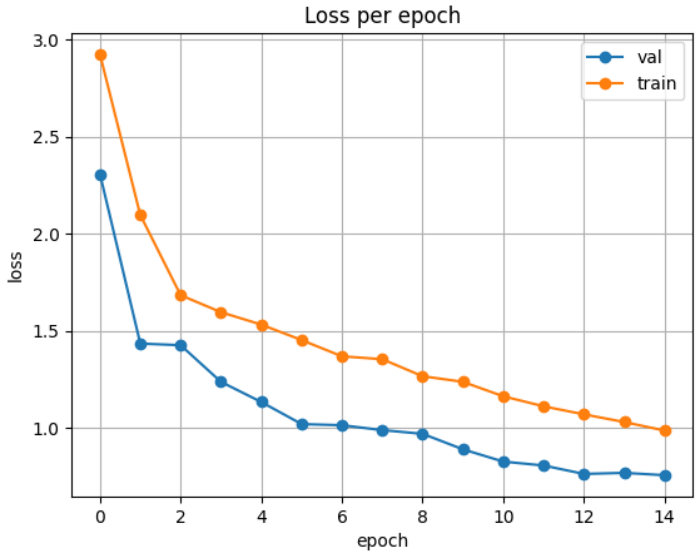

Figure 1. Iteration Loss Curve.

 doi:10.20944/preprints202508.1108.v1

Figure 1 illustrates the variation of training and validation losses over epochs. Initially, the loss values rapidly decrease, indicating rapid convergence of the model. Subsequently, the rate of reduction in loss decelerates and stabilizes, with the training loss eventually converging to approximately 0.8 and the validation loss stabilizing around 1.0.

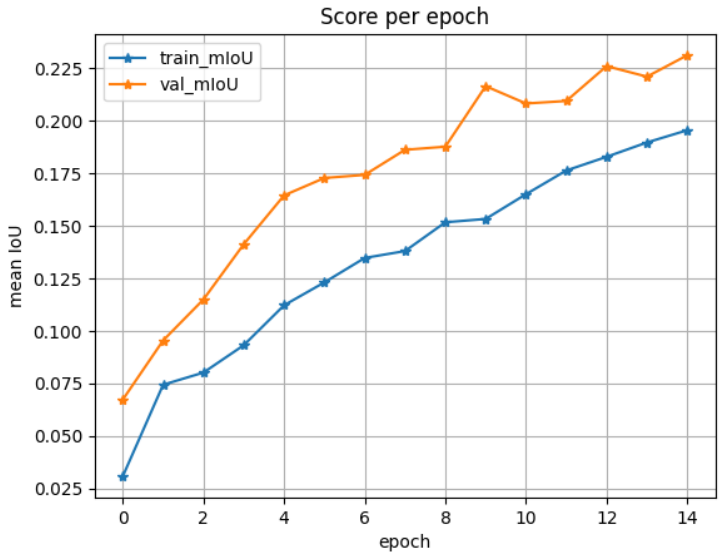

Figure 2. IoU Score Curve.

Figure 2 focuses on the core evaluation metric of the model, the mIoU (mean Intersection over Union), and its evolution during training. The plot clearly demonstrates that both the training and validation mIoU exhibit consistent and similar increasing trends as training progresses. This indicates the effectiveness of the model architecture and learning process, with steady improvement in the model's ability to accurately delineate boundaries of different classes. Crucially, the two curves show no significant divergence (e.g., premature decline or notably lower values in validation mIoU compared to training mIoU). The validation mIoU ultimately reaches approximately 0.2, which serves as a positive indicator of training stability, suggesting no evident overfitting or underfitting occurs.

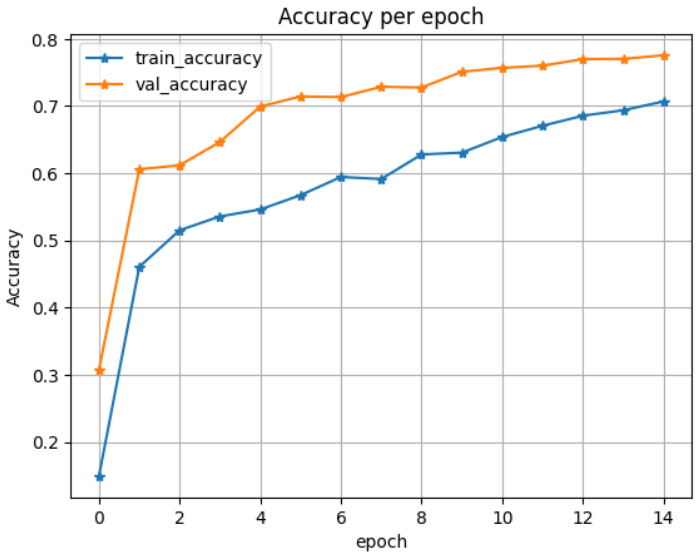

Figure 3. Accuracy Curve.

 doi:10.20944/preprints202508.1108.v1

Figure 3 illustrates the variation of overall accuracy in pixel-level classification. A notable observation is that the validation accuracy remains consistently stable and slightly higher than the training accuracy across most epochs, with both metrics ultimately converging to approximately 70%. This phenomenon, which is relatively uncommon, may suggest that the validation set samples are comparatively simple and exhibit minimal distributional differences from the training set. Indeed, this indicates that the model demonstrates robust generalization capability under the current data split.

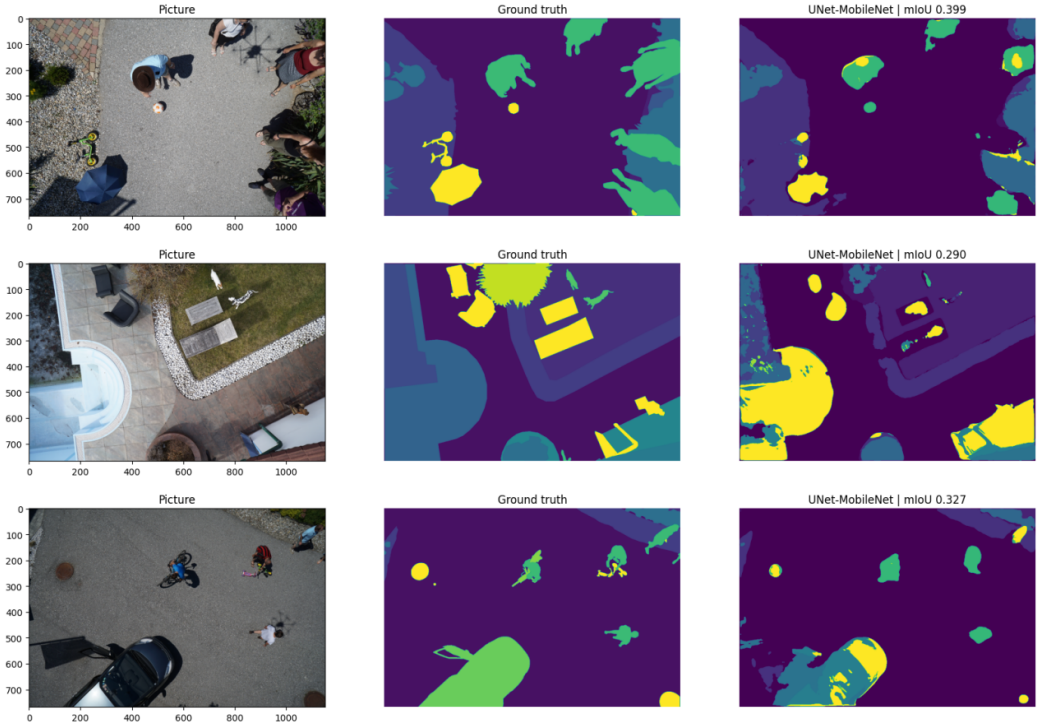

Figure 4. Visualization of Semantic Segmentation Results.

Figure 4 displays the semantic segmentation results on UAV-captured samples using the proposed U-Net + MobileNetV2 network. The outputs demonstrate that the model can correctly identify dominant objects such as roads, vegetation, and buildings, aligning with the ground truth labels in most test cases. However, segmentation accuracy near object boundaries remains suboptimal, and small targets are occasionally omitted or misclassified. These limitations may stem from the reduced capacity of the lightweight model, which trades off fine-grained spatial details for computational efficiency. Despite these challenges, the overall segmentation layout remains consistent with expected semantic regions, supporting the model's suitability for embedded deployment with further refinement.

Next, this section presents the performance optimization results achieved by adopting the RISC-V architecture. Custom instructions and FPGA acceleration significantly enhance computational efficiency. For standard convolution FLOPs F_{std} and depthwise separable convolution FLOPs F_{ds}, the speedup ratio is:

$$S = \frac{F_{std}}{F_{ds}} \approx \frac{C_{in}}{C_{out}} \tag{19}$$

For MobileNetV2, where $C_{in} \ll C_{out}$, the speedup ratio exceeds 10x.

On-chip SRAM reduces external memory access. For a tile size of $64 \times 64 \times C_{in}$, the data volume per tile is:

$$D = 64 \times 64 \times C_{in} \times 4 \tag{20}$$

 doi:10.20944/preprints202508.1108.v1

If the FPGA's local memory capacity is M_{fpga}, the number of required tiles is:

$$N = \left\lceil \frac{H \cdot W \cdot C_{in}}{64 \cdot 64 \cdot C_{in}} \right\rceil = \left\lceil \frac{H \cdot W}{4096} \right\rceil \tag{21}$$

The U-Net+MobileNetV2 model is deployed on the RISC-V heterogeneous processor. For an unoptimized model with inference time T_0, FPGA acceleration reduces it to T_1, yielding a speedup:

$$S_{time} = \frac{T_0}{T_1} \tag{22}$$

Adjusting tile size and pipeline depth further optimizes energy efficiency.

5. Conclusions

This work introduces a lightweight semantic-segmentation framework specifically designed for drones, where real-time inference and low power consumption are paramount. Leveraging a U-Net encoder-decoder with a MobileNetV2 backbone that uses depthwise-separable convolutions, the model achieves a strong balance between accuracy and computational efficiency. Custom RISC-V cores, FPGA accelerators, and targeted instruction-set extensions further accelerate convolution and activation operations. Experiments confirm high accuracy and competitive mIoU, while hardware acceleration delivers more than a ten-fold speed-up and cuts off-chip memory traffic to 1⁄4096 when the tile size is 64 × 64. Additional performance gains may be realized by refining tile dimensions and pipeline depth. Challenges remain in boundary delineation and small-object recognition. Future work will therefore explore (i) multimodal fusion with sensors such as LiDAR or thermal cameras, (ii) dynamic quantization for deeper compression, and (iii) industrial-grade environmental testing to validate system robustness. Overall, the study demonstrates that power consumption on drone platforms can be markedly reduced without sacrificing segmentation accuracy, charting a promising course for next-generation UAV semantic segmentation framework.

References

1. Zhou Yongjun, Luo Nan, Sun Yanchen, et al. Fine Segmentation Method for Apparent Multi-defect Image of Concrete Bridge[J/OL].Journal of Harbin Institute of Technology,1-14.
2. Du Sunwen, Song Ruiting, Gao Zhiyu, et al. Shadow extraction of UAV images in open-pit mine based on improved UNet3+[J/OL].Electronic Measurement Technology,1-11.
3. Shen Hao, Ge Quanbo, Wu Gaofeng. Unmanned aerial vehicle (UAV) sea segmentation algorithm based on shallow and deep features of backbone network[J/OL].Journal of Intelligent Systems,1-17.
4. Lee A ,Sagong S ,Yi K . LiDAR-based semantic segmentation of immobile elements via hierarchical transformer for urban autonomous driving[J]. Journal of Mechanical Science and Technology,2025,39(6):1-11.
5. Nasiri K ,Martin G W ,LaRocque D , et al. Using Citizen Science Data as Pre-Training for Semantic Segmentation of High-Resolution UAV Images for Natural Forests Post-Disturbance Assessment[J]. Forests,2025,16(4):616-616.
6. Han Jinchi, Wang Zhidong, Ma Hao, et al.Spike-FlexiCAS: RISC-V processor simulator that supports flexible configuration of cache architecture[J/OL].Journal of Software,1-16.
7. Xu Xuezheng, Yang Deheng, Wang Lu, et al.Proof of the same address order consistency theorem of RISC-V memory consistency model[J/OL].Journal of Software,1-19.

 doi:10.20944/preprints202508.1108.v1

8. HE Xuefei, PU Qingtao, WAN Jiawang, et al. Design of BLDC motor control system based on RISC-V architecture MCU[J].Industrial Control Computer,2025,38(04):143-145.
9. Banchelli F ,Jurado D ,Gasulla G M , et al. Exploring RISC-V long vector capabilities: A case study in Earth Sciences[J]. Future Generation Computer Systems,2026,174107932-107932.