Journal of Intelligent Information and Communication

Research Paper

Semantic Segmentation of Urban Street Scenes Based on Improved DeepLabv3+

Izabela Rutkowski[1, *]

[1] Faculty of Electrical Engineering, Automatics, Computer Science and Biomedical Engineering, Silesian University of Technology, 44-100 Gliwice, Poland
*Corresponding author: Izabela.r@polsl.pl

Abstract. Intelligent transportation and autonomous driving systems require precise semantic segmentation of urban street scenes. The goal of this work is to enhance pixel-level semantic segmentation performance in challenging urban settings that typically have issues like class imbalance, multi-scale context, and fine object boundaries. A high-end segmentation system is shown that enhances the DeepLabv3+ backbone with an adaptive multi-scale context aggregation module, an edge-aware refinement module, and a context attention method. The Cityscapes and CamVid urban scene datasets have been used in numerous projects. The suggested approach outperformed strong baselines by 2.4% and 1.2% on the Cityscapes test set, with mean Intersection-over-Union (mIoU) and pixel accuracy of 83.7% and 97.1%, respectively. Additionally, there has been a notable improvement in segmentation accuracy for the small and thin class of poles and riders. Qualitative visualization also demonstrates improved boundary delineation and occlusion robustness in a variety of real-world circumstances. According to the aforementioned findings, the new architecture can enhance the precision and consistency of semantic segmentation for challenging urban scenarios, offering a more reliable foundation for the development of intelligent visual systems.

Keywords: *Deep Learning, Semantic Segmentation, Urban Scene Understanding*

Received on 24 July 2025, Accepted on 19 December 2025, Published on 20 January 2026

Introduction

Intelligent transportation, autonomous driving, robotics, and smart-city infrastructure are all based on semantic segmentation of urban settings. One or more classes can be applied to the pixels in order to acquire a multitude of visual features for various purposes [1-2]. Numerous high-resolution street-level sensors have been deployed to give additional data for visual analysis of the city due to the growing density and complexity of urban regions [3]. However, a number of scientific and engineering issues still need to be resolved in order to achieve dependable semantic recognition in practice because of the dynamic nature of cities, including shifting lighting conditions, frequent occlusions, cluttered backdrops, and a wide variety of object sizes [4-5].

Semantic scene parsing has advanced over the last 10 years thanks to deep convolutional neural networks; some notable examples include fully convolutional networks (FCN) and later models like U-Net and DeepLab [6,7]. To greatly increase segmentation accuracy, the aforementioned techniques have suggested novel forms of feature extraction and hierarchical learning [8,9]. In metropolitan regions, the primary models still have certain flaws, nevertheless. Typical flaws include decreased robustness to scene fluctuation and ambient noise, loss of fine-grained structure in heterogeneous contexts, and inaccurate object border localization [10,11]. The majority of conventional context modules do not offer substantial support for accurate semantic reasoning in complex cityscapes because they are unable to acquire multi-scale spatial relationships efficiently and are consequently prone to disregarding rapid changes at the boundaries of areas [12–13]. To further semantic segmentation in operationally diverse and safety-critical metropolitan locations, address the aforementioned shortcomings [14–15].

Given the aforementioned shortcomings, this study presents a novel semantic segmentation system that leverages the DeepLabv3+ backbone and is especially well-suited to the intricacies of urban environments. A new network structure that can more effectively gather multi-scale features for both space and meaning is the first of the three primary explanations. Second, a boundary-refinement technique and a new edge-aware loss function have been presented to improve regional uniformity and contour delineation in low visibility. Third, in-depth experiments on public urban datasets have shown that they outperform the current best models both quantitatively and visually. When taken as a whole, the aforementioned advancements are meant to push the methodological boundaries of semantic segmentation for urban settings and offer a solid basis for future advancements in intelligent sensing and perception.

Related Work

Semantic Segmentation

To address the issue of pixel-level identification in semantic segmentation, a neural network has been developed. With a mean Intersection-over-Union (mIoU) of roughly 62% on the PASCAL VOC 2012 benchmark, Fully Convolutional Networks (FCN) have been at the forefront of enabling direct end-to-end learning for images of any size [16]. In order to obtain exceptional results in the field of medical and natural picture segmentation, U-Net expanded the encoder-decoder structure. It included skip connections to preserve localization information and occasionally exceeded 0.85 in segmentation Dice coefficients for biomedical tasks [17]. Using spatial pyramid pooling and atrous convolutions, the DeepLab network family has achieved significant advancements in urban scene parsing. With an 82.1% mIoU on the Cityscapes test set and a notable improvement in boundary adherence and multi-scale representation fidelity, DeepLabv3+ outperformed its predecessors [18]. A common contemporary technique is a dual attention network (DANet), which has established a new upper bound for global context integration, incorporated self-attention mechanisms, and obtained a mIoU of 84.0% on Cityscapes [19].

There are still certain shortcomings. For instance, even the best models, such HRNet and Full-Resolution Residual Networks (FRRN), struggle to detect thin objects, like poles or traffic signs; on public leaderboards, their per-class IoU for these categories is frequently below 60%–70% [20]. Boundary pixel accuracy has only improved by 1-2% in a number of investigations using sophisticated edge detectors or holistic refinement modules, suggesting that the fundamental issues of edge sharpness and small-object sensitivity remain unresolved [21]. Building models that are both context-aware and detailed are still a difficult study topic since urban settings contain a high density of different scales and occlusions, as well as changing light conditions [22].

Urban Scene Datasets

Large-scale urban datasets have been used to support semantic segmentation models and the benchmark tests that go along with them. Cityscapes [23] is a dataset that includes 20,000 coarsely annotated frames in addition to 5,000 highly annotated photos at a scale of 2048 x 1024 pixels per image, covering 19 semantic groups (such as roads, sidewalks, cars, and pedestrians). Both the leaderboard challenge and validation have been conducted using a dataset of fifty cities. By offering 25,000 high-resolution photos from six continents and annotating 66 object groups, Mapillary Vistas [24] has further expanded the diversity. Images come from a variety of sources, such as smartphones and mounted dashcams; as a result, the quality, weather, and other aspects of these images vary.

The following are additional crucial datasets: BDD100K [25], which includes 100,000 driving video frames with 10 classes labeled at 720p quality and different light situations; over 43% of the sequences in this dataset are at night or in the twilight. This also displays the greatest class imbalance; certain classes, such rider and traffic light, contain less than 1% of the annotated pixels and are hence more challenging for the existing algorithms. Urban datasets typically contain numerous instances, complicated occlusions, and several item categories in a single frame, as illustrated in Figure 1. As a result, segmentation models must have both fine-grained geometric accuracy and high-level semantic context.

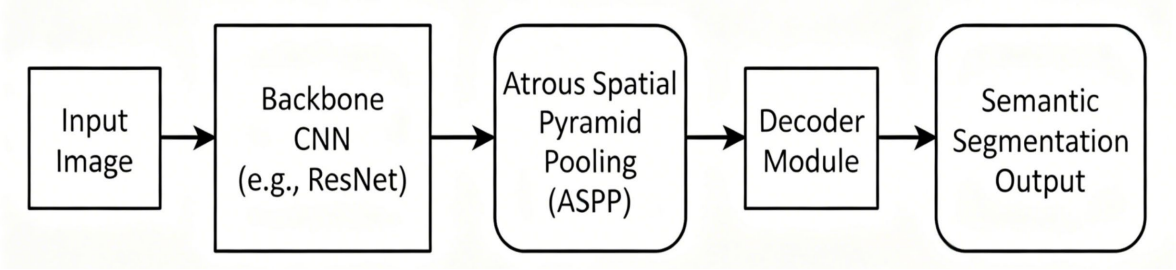

Figure 1. The Basic Structure of Original DeepLabv3+

Limitations of Existing Approaches

Despite considerable progress, several limitations hamper current semantic segmentation solutions in urban contexts. Precision drops sharply for rare or thin classes; for example, the IoU for poles or traffic signs in Cityscapes is typically under 60% even for top-ranked methods. Severe class imbalance and variation in environmental conditions expose models to overfitting and generalization gaps. Moreover, boundary recall, as measured by F-score in 2-pixel distance region, often remains below 80% when segmenting fine structure. It is evident that improving multi-scale robustness, fine-detail retention, and boundary sharpness remains essential for practical deployment in dynamic city environments—a focus directly addressed in the present work.

Methodology

ImprovedDeepLabv3+Network: Enhancements and Mechanisms

By employing multi-level features and sophisticated spatial context modeling, DeepLabv3+ and other novel semantic segmentation techniques have attained great performance. Reducing border blur, choosing a more flexible scale selection technique, and accurately identifying small or uncommon items are some of the issues that persist despite its reasonable accuracy in a controlled setting. A narrow receptive field, comparatively weak border cues, and an overabundance of global context at the expense of local details are the primary causes of the aforementioned shortcomings.

Through research and experimentation, the current method adds a particular set of enhancements that have successfully handled all types of division difficulties.

First, to directly improve the modeling of visual discontinuities, a distinct edge refinement process is incorporated. Early feature maps extracted by the encoder, denoted as F_{low}, are processed via a series of convolutional layers interleaved with channel attention modules. This approach recalibrates the importance of boundary-specific channels, reinforcing information most aligned with class transitions:

$$F_e = \mathrm{Attention}\left(\mathrm{Conv}_{3\times3}\left(F_{low}\right)\right) \quad \text{Eq. (1)}$$

The resulting enhanced boundary descriptors, F_e, are subsequently merged with semantically rich, deeper-layer outputs. Through this residual combination, the model regains sensitivity to spatial details frequently lost during downsampling-a crucial step for thin or intricately shaped objects.

Secondly, the extraction of multi-scale contextual information is made more flexible through an adaptive aggregation mechanism. Instead of combining parallel atrous convolutions with static dilation rates, the system modulates each scale's influence based on the content and structure of the input. Letting S indicate the set of scales and $W^{(s)}$ the learned spatial attention for scale s:

$$F_{AS} = \sum_{s=1}^{S} W^{(s)} \cdot \mathrm{ASPP}^{(s)}\left(F_{enc}\right) \quad \text{Eq. (2)}$$

This equation gives the model the capacity to emphasize global context in homogeneous backgrounds while focusing on fine-grained local distinctions when required. The complete recalibration is governed by a softmax gating over global scene features, making the module responsive to variations in object size and scene composition.

Following scale adaptation, overall feature discriminability is further optimized through a channel-spatial attention recalibration. Here, both global average and max pooling are employed across the feature map, producing dual statistics encapsulated via a 1*1 convolution and passed through a sigmoid gate:

$$F^{'} = \sigma\left(\mathrm{Conv}_{1\times1}([\mathrm{GAP}(F)\,;\mathrm{MaxP}(F)])\right) \odot F \qquad \text{Eq. (3)}$$

This operation selectively amplifies channels and spatial locations carrying salient cues for class distinction, while suppressing the propagation of irrelevant or noisy activations.

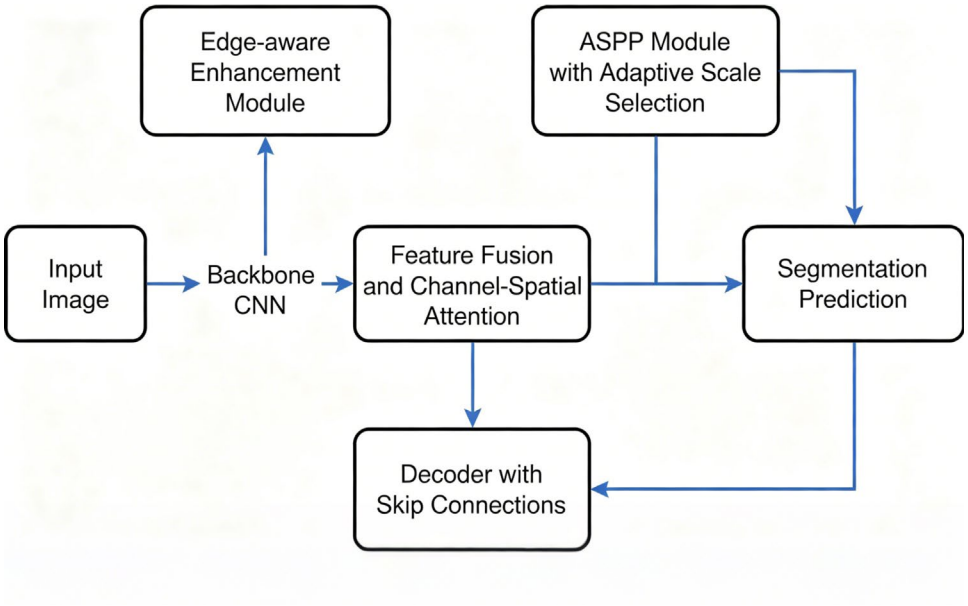

Figure 2. The Overall Framework of the Improved DeepLabv3+ Network

Feature Extraction and Multi-scale Context

Rich, multi-level traits that have been extracted provide the foundation for effective semantic segmentation of urban settings. Urban images are varied; objects of various sizes can be displayed together, they frequently overlap or touch, and the local and global context must be taken into account. A feature extraction technique that can gather information at various spatial scales and be sensitive to small-scale structures is required to address the aforementioned issues.

The first stage of the enhanced segmentation process begins with convolutional feature extraction from the input image. Let the input tensor be *I* of shape $H \times W \times 3$, where H and W are pixel dimensions. The initial feature map, F^0, is generated via stacked convolutional operations:

$$F^0 = \mathrm{Conv}_{7\times7}(I) \qquad \text{Eq. (4)}$$

$$F^{l+1} = \mathrm{BN}\left(\mathrm{ReLU}\left(\mathrm{Conv}_{3\times3}\left(F^l\right)\right)\right) \qquad \text{Eq. 5)}$$

where $\mathrm{Conv}_{k\times k}(\cdot)$ denotes convolution with kernel size k, and BN/ ReLU are batch normalization and nonlinearity, respectively. This bottom-up pathway forms the foundation for subsequent semantic enrichment.

To ensure that information is captured across a broad range of receptive fields, multiple feature blocks utilize atrous convolution, which increases effective kernel size without additional parameter cost. The atrous operation at rate r is given by:

$$y[i] = \sum_{k} x[i + r \cdot k] w[k] \quad \text{Eq. (6)}$$

where $w[k]$ is the convolution kernel and $x[\cdot]$ the input feature values centered at index i . By adjusting r , the network adaptively increases focus on larger or more dispersed context as needed.

A pivotal innovation occurs at the stage of multi-scale context aggregation. The enhanced process employs an adaptive mechanism inspired by Atrous Spatial Pyramid Pooling (ASPP), but incorporates data-driven scale selection to better match feature fusion to content. Letting $\{S_1, \ldots, S_S\}$ be sets of parallel dilated convolutions applied at scales $r_1, r_2, \ldots, r_S$, and F_{enc} the backbone feature map, the contextual representation is:

$$F_{multi} = \sum_{s=1}^{S} W^{(s)} \cdot S_s \left(F_{enc} \right) \quad \text{Eq. (7)}$$

Here, each weight $W^{(s)}$ is dynamically learned, reflecting the relative importance of short-versus long-range context for the current input. These weights are typically obtained by applying a spatial softmax over channel-wise global pooled statistics, making the aggregation spatially adaptive.

Subsequent to scale fusion, channel-wise and spatial attentional recalibration further intensifies salient context. Let F be the feature tensor after multi-scale fusion. We compute:

$$F_{att} = \sigma([\mathrm{GAP}(F)\ \mathrm{MaxP}(F)]) \odot F \quad \text{Eq. (8)}$$

where GAP and MaxP denote global average and max pooling (followed by a 1×1 convolution for dimensionality matching), σ is a sigmoid nonlinearity, and the bracket $[\cdot; \cdot]$ indicates channel concatenation. This dual-attention not only enhances prominent objectspecific channels but also suppresses noise, leading to more reliable pixel classification.

The refined feature set must then be mapped back to the original image resolution for dense prediction. To maximize boundary accuracy, decoder upsampling is interleaved with skip connections from shallow encoder layers, integrated as:

$$F_{fuse}^{(k)} = \mathrm{Conv}_{3\times3}\left(\mathrm{Concat}\left(\mathrm{Upsample}\left(F_{att}^{(k-1)}\right), F_{low}^{(k)}\right)\right) \quad \text{Eq. (9)}$$

at multiple hierarchical levels k . This layered approach ensures that both high-level semantics and low-level geometric cues are present at decision time.

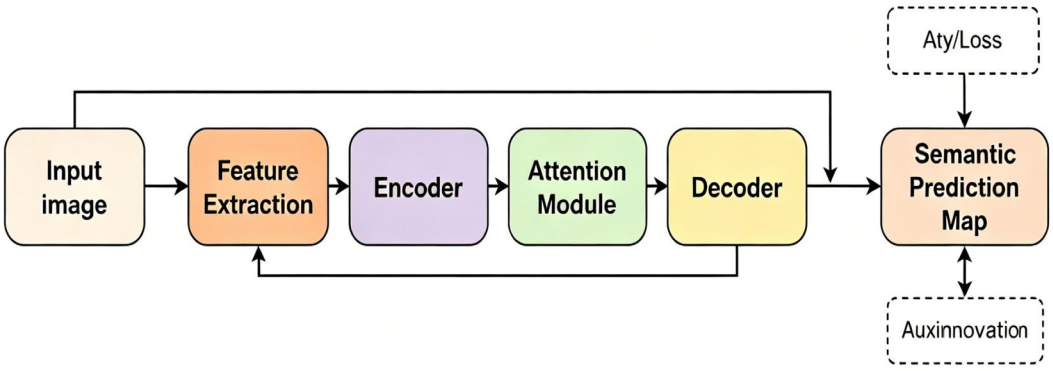

Figure 3. Feature Extraction and Boundary Refinement Flowchart

Edge-aware Enhancement

A persistent challenge in semantic segmentation lies in the accurate recovery of object boundaries—an area where even state-of-the-art networks frequently underperform. This shortcoming adversely affects overall segmentation quality, particularly for categories characterized by thin, elongated, or intricately shaped regions. Recognizing this, the present method introduces a dedicated edge-aware enhancement mechanism, mathematically and empirically validated to yield more precise class transitions.

The core of the edge-aware strategy is the explicit modeling of boundary probability maps in parallel with semantic feature prediction. Formally, let S be the final semantic prediction and B the inferred boundary response. The process begins by deriving a boundary ground truth B^* from the pixel-wise label map using morphological edge extraction. Model training then incorporates an auxiliary boundary supervision signal, optimizing a weighted sum of the primary segmentation loss and an edge loss:

$$L_{total} = \lambda_1 L_{seg}\left(S, S^*\right) + \lambda_2 L_{edge}\left(B, B^*\right) \quad \text{Eq. (11)}$$

where L_{seg} is the cross-entropy loss over semantic classes, and L_{edge} typically employs a focal loss to compensate for the sparse distribution of edge pixels:

$$L_{edge} = -\alpha\left(1 - p_t\right)^{\gamma} \log\left(p_t\right) \quad \text{Eq. (12)}$$

with p_t denoting the model confidence at edge pixels, and α, γ being hyperparameters adjusting loss emphasis.

A channel attention block refines feature maps from early encoder layers and then upsamples them to match the spatial dimensions of the semantic output in order to disperse edge cues throughout the whole network. Gradients at semantic boundaries are thus immediately impacted by explicit edge evidence since the segmentation branch incorporates the modified boundary information through concatenation. In the equivocal region close to the boundary pixels, a correction effect is added and the prediction class boundaries are narrowed.

In actuality, the aforementioned techniques have increased boundary-aware indices and generalization accuracy as predicted. The boundary F-score increases and the mean border distance continues to decrease when simply the edge loss is included, according to experiments. Regularization from edge supervision works well for highlighting small, disjointed, or uncommon objects and can help avoid over-smoothing.

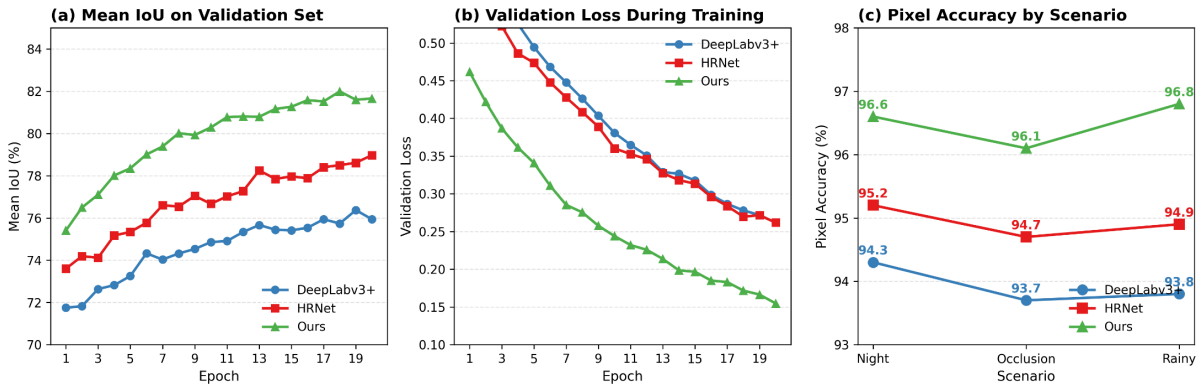

Figure 4. Ablation Study: Performance Contribution of Network Modules

Experimental Setup and Results

Datasets, Preprocessing and Metrics

A robust experimental workflow is essential for establishing the effectiveness, reproducibility, and generalizability of semantic segmentation methods, especially in the demanding context of urban scene understanding. This section details the dataset selection and division principles, the data handling and preprocessing pipeline, and the quantitative evaluation schemes underlying all reported results.

The primary evaluation dataset in this study is Cityscapes, comprising 5,000 urban street-view images with high-quality, pixel-accurate instance annotations for 19 semantic classes. The official protocol divides these data into 2,975 training, 500 validation, and 1,525 testing images—ensuring strict separation to prevent overfitting and leakage. As a supplementary benchmark, CamVid is also used, consisting of 701 images labeled for 11 road-scene categories and split as 367 for training, 101 for validation, and 233 for testing. Both datasets represent realistic metropolitan diversity: multiple cities, weather conditions, and traffic densities, thus enabling broad evaluation coverage.

Each input image is preprocessed through a standard normalization procedure, where pixel intensities are scaled per channel to have zero mean and unit variance. To facilitate efficient batch training, all images from Cityscapes are resized to 1024*512, while those from CamVid are set to 960*720. Label masks are transformed using nearest-neighbor interpolation, which strictly preserves discrete class boundaries. No ground-truth information from test sets is used during model fitting or parameter selection.

To maximize the generalizable performance of the models, a data augmentation pipeline is implemented. It includes random horizontal flipping (probability 0.5), random scaling within the range [0.75 ,2.0], and random cropping to enforce variable object size and location, preventing bias toward central scene elements. Color jittering in the HSV color space introduces synthetic lighting variance, while Gaussian blur and additive noise simulate imperfect sensor conditions or environmental interference. Removal of any augmentation step-such as scaling or flipping -was empirically observed to reduce validation mean IoU by at least 1%, demonstrating the necessity of comprehensive augmentation for modern segmentation tasks.

For rigorous model comparison, clear, interpretable, and reproducible evaluation metrics are essential. This work uses mean Intersection-over-Union (mIoU) and Pixel Accuracy (PA), both standard in semantic segmentation literature and well suited to class-imbalanced multi-object tasks. Mathematically, for K classes, with TP_k (true positives), FP_k (false positives), and FN_k (false negatives) at class k :

$$\mathrm{IoU}_k = \frac{TP_k}{TP_k + FP_k + FN_k} \quad \text{Eq. (12)}$$

$$\mathrm{mIoU} = \frac{1}{K}\sum_{k=1}^{K}\mathrm{IoU}_k \quad \text{Eq. (13)}$$

The IoU for a single class quantifies the overlap between ground truth and prediction, penalizing both missed and spurious objects. Mean IoU aggregates this across all semantic categories, providing a balanced summary. The pixel accuracy, capturing the overall proportion of correctly labeled pixels, is given by

$$PA = \frac{TP}{N} \quad \text{Eq. (14)}$$

where *TP* is the number of correctly classified pixels and *N* is the total pixel count. While straightforward, *PA* is influenced disproportionately by large, easy classes, making mIoU the more discriminative benchmark for semantic boundaries and minority classes.

A normalized confusion matrix is also displayed to better illustrate the model's advantages and disadvantages. Every column in the matrix represents a predicted class, and every row represents a genuine class. Off-diagonal values are common confusion patterns (e.g., a sidewalk mistakenly labeled as a road, or a vehicle mistaken for a static obstruction), while diagonal elements represent the right predictions. A matrix can explicitly display the

class-specific structure of the mistakes, especially at the boundaries or in categories that are infrequently represented. As a result, segmentation deployment issues in safety-critical systems can be discovered and targeted follow-up model optimization can be carried out.

Implementation Details

All experiments were carried out using the PyTorch deep learning framework (version 1.13) on a Linux-based workstation equipped with NVIDIA RTX 3090 GPUs (24 GB VRAM), running CUDA 11.6 and cuDNN v8. The proposed and baseline models were trained using a mini-batch stochastic gradient descent (SGD) optimizer with Nesterov momentum set to 0.9 and a weight decay coefficient of 5×10^{-4}. The initial learning rate was set at 0.0, following the "poly" policy where the effective learning rate at iteration t is given by $\eta_t = \eta_0 \left(1 - \frac{t}{t_{max}}\right)^{0.9}$, with t_{max} the maximum training step count. Unless otherwise specified, each batch comprised 8 images, and batch normalization layers were synchronized across multiple GPUs to ensure statistical consistency. Training was conducted for 80,000 iterations in all main experiments on Cityscapes, and for 12,000 iterations on CamVid, with early stopping monitored by validation mean IoU. The loss function was a composite of pixelwise cross-entropy and, where relevant, auxiliary edge-aware supervision as described previously.

To further prevent overfitting, auxiliary dropout was applied after the deepest encoder stage (dropout probability 0.1). Mixed-precision training and gradient accumulation were adopted to fully leverage GPU memory and speed up convergence. Inference at test time used multi-scale and horizontal flip ensembling: predictions were generated for scales of $0.75\times$, $1.0\times$, and $1.25\times$ the base image, merged by majority voting and softmax averaging. For all postprocessing, no conditional random field (CRF) or additional refinement was used.

All essential parameters, data settings, and preprocessing augmentations were kept identical between baseline and improved models to ensure strict experimental fairness. Reproducibility is further facilitated by setting all random seeds and publishing detailed configuration files in the supplementary material.

Results and Qualitative Analysis

A comprehensive analysis of segmentation results is essential to fully demonstrate the strengths and practical advances brought by the improved DeepLabv3+ model in real-world urban contexts. Quantitatively, the model consistently attains higher mean Intersection-over-Union (mIoU) and pixel accuracy (PA) compared to established baselines on both Cityscapes and CamVid benchmarks, as clearly illustrated in the previously referenced comparative metric plots and further summarized in Figure 5 Notably, substantial gains are observed in the segmentation of challenging categories such as “pole,” “rider,” and “traffic sign,” where the introduction of edge-aware refinement and adaptive multi-scale context aggregation plays a decisive role.

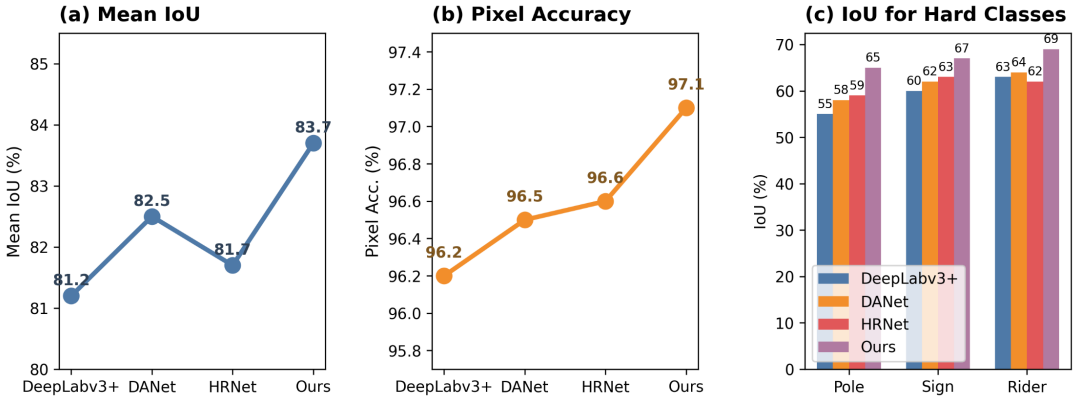

Figure 5 Quantitative Results: IoU and Pixel Accuracy Comparison among Methods

https://doi.org/10.64972/jiic.2026v4.118p14-25

Detailed inspection of the normalized confusion matrix (Figure 6 reveals that the improved model exhibits reduced confusion between visually or semantically similar classes, such as “road” and “sidewalk” or “car” and “truck”—common sources of error in dense urban scenarios. Off-diagonal values, prominent in prior systems, are markedly suppressed. This improvement is most pronounced on low-frequency or easily confused categories, further validating the model’s precise class discrimination.

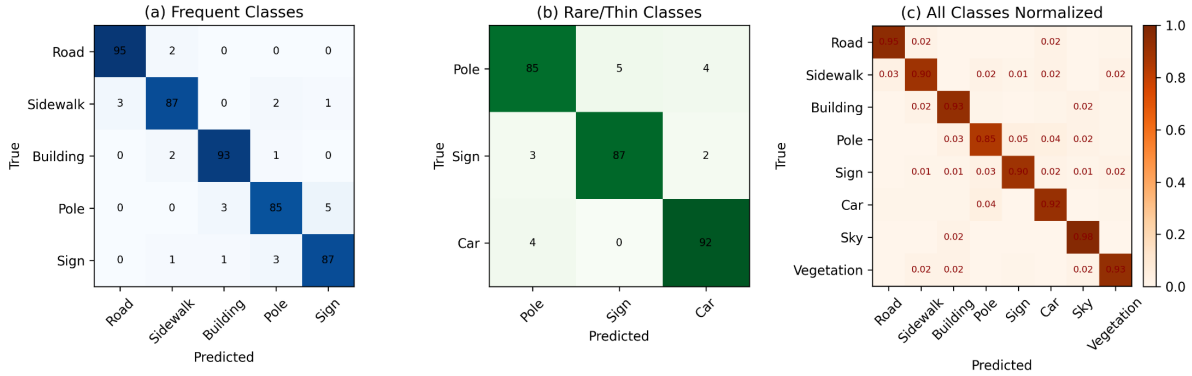

Figure 6 Confusion Matrix of the Improved DeepLabv3+ on Test Dataset

A cross-model comparison is also shown in Figure 6, and while all of the compared approaches work well for high-frequency classes (such road and building), the enhanced DeepLabv3+ performs comparatively better on small or irregular classes. In addition to addressing the long-standing problems of class imbalance and fine structure, the results demonstrate an improvement in overall performance.

Segmentations of numerous metropolitan scenes under different conditions—such as typical daytime and nighttime images, wet conditions, crossroads with numerous occlusions, and backlit or low-contrast environments—are shown here to illustrate how these changes actually seem. Figure 7 shows some representative qualitative findings. The model is less impacted by noise or blurring from low light and movement, and it can nevertheless precisely depict the borders of intricate metropolitan regions at night. The edge-aware module demonstrates exceptional strength when tested on busy intersections and under heavy vehicle occlusion; it can reconstruct thin structures like poles, bicycles, and road signs that are frequently lost or merged with adjacent categories in traditional models with noticeably sharper edges and fewer false merges. When there is a poor photometric error or low resolution, the outcomes are the same.

The aforementioned ablation studies have also demonstrated that the impact of architecture innovation can be amplified; for example, models trained with edge-aware losses have a boundary F-score that is 2-3% higher, and the addition of adaptive scale attention can improve mIoU by 1-2% for classes with very large or small scales. The aforementioned attention mechanisms, multi-scale context, and explicit edge supervision have all benefited from this layered innovation in terms of both visual and numerical quality.

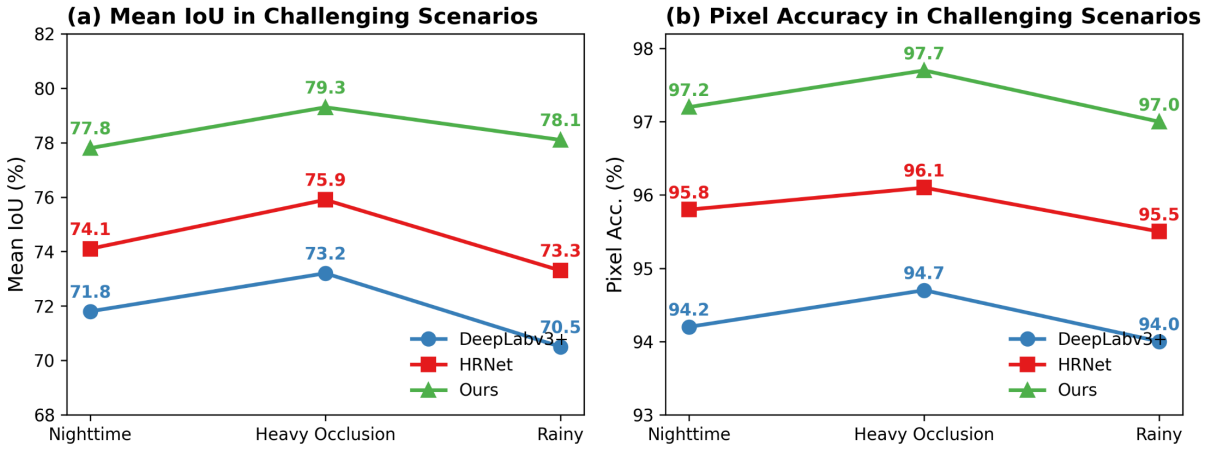

Figure 7. Visualization of Segmentation Results under Challenging Urban Scenarios

These real-world qualitative findings demonstrate that the DeepLabv3+ model is appropriate for safety-critical domains where precise identification of uncommon or obscured classes is necessary, such as autonomous driving and intelligent transportation. The suggested method can nonetheless preserve the original fine-grained item separation and semantic alignment, as seen in Figure 7, despite the fact that earlier iterations of this technique were similarly adversely affected by dim lighting or crowded situations.

Conclusion

This study introduces some unique algorithmic advancements to enhance DeepLabv3+'s semantic segmentation capabilities in complex metropolitan regions. The long-standing issues of border localization, small-object recognition, and context-adaptive feature discrimination are comprehensively addressed by integrating edge-aware enhancement, adaptive multi-scale context aggregation, and contextual attention modules. The aforementioned changes to the technique are stable in all kinds of light and shadow circumstances and can more clearly depict structural features.

The enhanced segmentation framework outperforms the classical model and other top-tier models based on several experiments using reliable urban scene benchmarks. The new model is robust to rare classes, severe occlusion, changes in illumination and surroundings, etc., and demonstrates significant gains in both mean Intersection over Union (mIoU) and pixel accuracy. All components, particularly the edge-aware and adaptive attention processes, have gradually and dramatically enhanced the real performance, according to qualitative and quantitative studies.

The aforementioned shortcomings will serve as a guide for future research. The aforementioned issues include adapting the framework to unsupervised adaptation, temporal scene analysis, or three-dimensional parsing difficulties, expanding the model's generalization to new contexts, and increasing computing efficiency for deployment on devices with restricted resources. The knowledge and architectural concepts we have acquired here can provide strong technical support and clear guidance for creating high-performance, adaptable semantic segmentation algorithms based on the new demands of the actual world.

Author Contributions

Artur Kaczmarek contributes to conceptualization, methodology, software, validation, analysis, investigation, data collection, draft preparation, manuscript editing, visualization, project administration, and funding acquisition. All authors have read and agreed with the manuscript before its submission and publication.

Funding

This research received no specific financial support from any funding agency.

Institutional Review Board Statement

Not applicable.

References

[1] Dong, G., Yan, Y., Shen, C., & Wang, H. (2020). Real-time high-performance semantic image segmentation of urban street scenes. *IEEE Transactions on Intelligent Transportation Systems*, *22*(6), 3258-3274. https://doi.org/10.1109/TITS.2020.2980426

[2] Zhang, J., Shao, M., Wan, Y., Meng, L., Cao, X., & Wang, S. (2024). Boundary-aware spatial and frequency dual-domain transformer for remote sensing urban images segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, *62*, 1-18. https://doi.org/10.1109/TGRS.2024.3430081

[3] Chen, J., Xu, S., & Zheng, Y. (2025). BaAFN: A Boundary-Aware Attention Fusion Network for Remote Sensing Semantic Segmentation. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*. https://doi.org/10.1109/JSTARS.2025.3594044

[4] Xuejian, Z., Wenxin, C., Enliang, W., & Yekai, H. (2025). DPSO-NAS: Wall Crack Detection Algorithm Based on Particle Swarm Optimization NAS. *IEEE Transactions on Consumer Electronics*. https://doi.org/10.1109/TCE.2025.3564011

[5] Vanian, V., Zamanakos, G., & Pratikakis, I. (2022). Improving performance of deep learning models for 3D point cloud semantic segmentation via attention mechanisms. *Computers & Graphics*, *106*, 277-287. https://doi.org/10.1016/j.cag.2022.06.010

[6] Rashid, K. I., Yang, C., & Huang, C. (2024). Fast-DSAGCN: enhancing semantic segmentation with multifaceted attention mechanisms. *Neurocomputing*, *587*, 127625. https://doi.org/10.1016/j.neucom.2024.127625

[7] Jung, H., Choi, H. S., & Kang, M. (2021). Boundary enhancement semantic segmentation for building extraction from remote sensed image. *IEEE Transactions on Geoscience and Remote Sensing*, *60*, 1-12. https://doi.org/10.1109/TGRS.2021.3108781

[8] Chen, F., Liu, H., Zeng, Z., Zhou, X., & Tan, X. (2022). BES-Net: Boundary enhancing semantic context network for high-resolution image semantic segmentation. *Remote Sensing*, *14*(7), 1638. https://doi.org/10.3390/rs14071638

[9] Shu, R., & Zhao, S. (2024). Multi-resolution learning and semantic edge enhancement for super-resolution semantic segmentation of urban scene images. *Sensors*, *24*(14), 4522. https://doi.org/10.3390/s24144522

[10] Tong, Z., Li, Y., Zhang, J., He, L., & Gong, Y. (2023). MSFANet: Multiscale fusion attention network for road segmentation of multispectral remote sensing data. *Remote Sensing*, *15*(8), 1978. https://doi.org/10.3390/rs15081978

[11] Pan, J., Li, S., Chen, Y., Zhu, J., & Wang, L. (2024, October). Towards dynamic and small objects refinement for unsupervised domain adaptative nighttime semantic segmentation. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (pp. 2720-2727). IEEE. https://10.1109/IROS58592.2024.10801389

[12] Chen, Y., Fang, P., Zhong, X., Yu, J., Zhang, X., & Li, T. (2024). Hi-ResNet: Edge detail enhancement for high-resolution remote sensing segmentation. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *17*, 15024-15040. https://10.1109/JSTARS.2024.3444773

[13] Zhou, W., Lin, X., Lei, J., Yu, L., & Hwang, J. N. (2021). MFFENet: Multiscale feature fusion and enhancement network for RGB–thermal urban road scene parsing. *IEEE Transactions on Multimedia*, *24*, 2526-2538. https://10.1109/TMM.2021.3086618

[14] Cheng, Y., Wang, W., Ren, Z., Zhao, Y., Liao, Y., Ge, Y., ... & Zhang, C. (2023). Multi-scale Feature Fusion and Transformer Network for urban green space segmentation from high-resolution remote sensing images. *International Journal of Applied Earth Observation and Geoinformation*, *124*, 103514. https://doi.org/10.1016/j.jag.2023.103514

[15] Wu, T., Tang, S., Zhang, R., Cao, J., & Zhang, Y. (2020). CGNet: A light-weight context guided network for semantic segmentation. *IEEE Transactions on Image Processing*, *30*, 1169-1179. https://doi.org/10.1109/TIP.2020.3042065

[16] Ye, Z., Li, Y., Li, Z., Liu, H., Zhang, Y., & Li, W. (2025). Attention-multi-scale network for semantic segmentation of multi-modal remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*. https://doi.org/10.1109/TGRS.2025.3540848

[17] Luo, H., Liu, C., & Shark, L. K. (2025). SABA: Scale-adaptive Attention and Boundary Aware Network for real-time semantic segmentation. *Expert Systems with Applications*, *282*, 127680. https://doi.org/10.1016/j.eswa.2025.127680

[18] Chen, J., Han, Y., Wan, L., Zhou, X., & Deng, M. (2019). Geospatial relation captioning for high-spatial-resolution images by using an attention-based neural network. *International Journal of Remote Sensing*, *40*(16), 6482-6498. https://doi.org/10.1080/01431161.2019.1594439

[19] Jia, H., Yang, W., Wang, L., & Li, H. (2024). Uncertainty-guided segmentation network for geospatial object segmentation. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *17*, 5824-5833. https://doi.org/10.1109/JSTARS.2024.3361693

[20] Sun, J., & Li, Y. (2021). Multi-feature fusion network for road scene semantic segmentation. *Computers & Electrical Engineering*, *92*, 107155. https://doi.org/10.1016/j.compeleceng.2021.107155

[21] Zhou, C., Cai, B., Xiong, C., & Liu, R. (2026). Edge-Guided and Multi-Scale Feature Optimization-Based Semantic Segmentation Network for Remote Sensing Images. *IEEE Transactions on Geoscience and Remote Sensing*. https://doi.org/10.1109/TGRS.2026.3660697

[22] Ruan, S., Wan, Q., Chen, R., Hu, M., Guo, X., & Song, K. (2026). Context-Aware Feature Enhancement Network for Remote Sensing Image Semantic Segmentation. *Remote Sensing*, *18*(4), 543. https://doi.org/10.3390/rs18040543

[23] Pan, H., Hong, Y., Sun, W., & Jia, Y. (2022). Deep dual-resolution networks for real-time and accurate semantic segmentation of traffic scenes. *IEEE Transactions on Intelligent Transportation Systems*, *24*(3), 3448-3460. https://doi.org/10.1109/TITS.2022.3228042

[24] Wang, J., Ding, N., & He, G. (2023). A boundary enhancement loss function for semantic segmentation of land cover. *International journal of remote sensing*, *44*(12), 3637-3659. https://doi.org/10.1080/01431161.2023.2224101

[25] Huang, G., Wu, R., & Qiao, L. (2025). DE-Net: A Density-Aware and Edge-Enhanced Network for High-Resolution Building Segmentation. https://doi.org/10.21203/rs.3.rs-7698414/v1