Journal of Applied Automation Technologies

Research Paper

Parallax-Aware Urban Road Scene Segmentation Based on 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. In order to promote the development of intelligent transportation and autonomous driving systems, clear semantic segmentation is essential. Due to depth discontinuities and parallax artifacts from dynamic viewpoints, as well as spatial misalignment and segmentation errors, accurately analyzing complex urban scenes is very challenging. This paper proposes a complete urban road segmentation framework. By introducing a separate disparity removal module in DeepLabv3+, the issue of disparity effects is clearly addressed. Model and correct the geometric distortions caused by depth, and then perform multi-scale semantic feature fusion for detailed scene analysis. All standard public datasets contain different levels of annotation details, including various urban areas and environments. According to the experimental results, the proposed method significantly improves the mean Intersection over Union (IoU) and boundary accuracy under conditions of strong perspective distortion, occlusion, and multi-layer structures. These conditions are contrary to the current baseline model. The architecture is very reliable under various weather conditions and urban environments, and it can run in real-time on embedded systems. Some improvements have been made, but there are still issues under extremely low visibility conditions. In the future, by integrating multimodal sensors and enhancing data augmentation, it is expected to provide more solutions. This study proposes a disparity-aware segmentation method for urban road scenes, directly supporting smart mobility and smart infrastructure.

Keywords: *Urban Computing, Semantic Segmentation, Urban Scene Understanding, Parallax Correction, DeepLabv3+, Intelligent Transportation*

Received on 19 July 2025, Accepted on 13 December 2025, Published on 19 January 2026

Copyright © 2026 Author, licensed to JAAT. This is an open access article distributed under the terms of the CC BY-NC-SA 4.0, which permits copying, redistributing, remixing, transformation, and building upon the material in any medium so long as the original work is properly cited.

Introduction

With the development of intelligent transportation systems, higher demands are being placed on the accuracy and reliability of semantic perception of urban road environments. A good division of urban areas will provide a technical foundation for autonomous vehicles, intelligent driving, advanced driver-assistance systems (ADAS), and other smart city applications, enabling traffic analysis, infrastructure monitoring, and real-time navigation [1]. With the widespread installation of cameras and the proliferation of connected vehicles, the variety and quantity of urban visual data are continuously increasing. These changes bring new opportunities and challenges for automated perception systems. High-quality semantic segmentation can not only support digital maps, urban planning, and smart infrastructure analysis, but also help achieve safe driving and situational awareness [2,3]. Due to the continuous growth of population density, density, and urban scale [4,5], the accuracy and reliability of these systems are crucial to ensure public safety and the efficient development of urban traffic [6]. Lighting fluctuations, frequent occlusions, and the constant changes of moving and stationary vehicles on the road necessitate segmentation algorithms with broad generalization capabilities [7,8].

Due to the non-planar city geometry and depth discontinuities, traditional computer vision and conventional learning methods often struggle to handle parallax artifacts [9]. Some progress has been made, but this problem still persists. Due to the fact that the aforementioned common methods usually assume local planarity or use

https://doi.org/10.64972/jaat.2026v4.109

shallow feature representations, segmentation boundary distortion or object misclassification is a common phenomenon in cases of overlapping structures, multiple camera perspectives, or multi-layered road layouts [10, 11]. According to recent benchmarks [12], the ability to detect lanes, estimate drivable areas, and understand scenes may be affected by parallax, which can reduce semantic accuracy. Recent research has progressed toward geometric models and data-driven adaptation, but a complete method for achieving stable disparity correction and high-definition segmentation has not yet been found.

This paper introduces a new system that combines explicit disparity removal with the DeepLabv3+ segmentation backbone network, based on advancements in deep learning. Dynamic modeling and correction of depth-based geometric distortion are employed to enhance the semantic accuracy and boundary clarity of complex urban scenes that traditional pipelines cannot handle. Designing a disparity-aware correction module, end-to-end geometric-semantic fusion in segmentation networks, and comprehensive validation using multiple real-world datasets are all contributions.

Literature Review

Urban Scene Segmentation Approaches

The development of intelligent transportation systems now requires the semantic segmentation of urban road scenes. Classic machine learning algorithms in the early stages of urban applications relied on handcrafted features, such as edges, color, and texture descriptors, to distinguish between road elements and the background [13]. To extend the aforementioned methods for boundary localization and context awareness, probabilistic graphical models, such as Conditional Random Fields [14], are used. Due to the increasing scale and complexity of urban scenes, the original model can no longer adapt to the various lighting conditions, obstacles, and objects in these dynamic street scene graphs [15].

Hierarchical representation learning enables convolutional neural networks to distinguish small-scale semantic categories. Fully Convolutional Networks enable end-to-end training on large-scale data, addressing the localization issues based on patches or sliding windows [16]. By introducing multi-scale feature aggregation, spatial attention, and context modules, the issues of complex structural urban areas and multi-scale features were addressed. The model has also been modified to meet the increasing robustness, efficiency, and cross-domain generalization requirements of all benchmark sites [17]. The aforementioned changes have laid the foundation for semantic segmentation in urban areas, but there are still some fundamental issues, such as geometric inaccuracies and spatial inaccuracies.

Parallax Correction Methods

Due to depth discontinuities and changes in camera position, parallax artifacts may cause geometric alignment issues in semantic segmentation results of urban areas. In the past, geometric methods were used to address these artifacts. By using homography or depth information in stereo images, other non-planar street objects have also been undistorted [18]. The aforementioned solutions are theoretically reasonable, but due to the requirements for precise feature matching and the relative rigidity of their geometric assumptions, they perform poorly in unconstrained or rapidly changing environments [19].

With the increasing availability of depth sensors and improvements in structured light technology, many new processes that directly introduce auxiliary geometric cues during the segmentation phase have begun. Due to the alignment of scene representations and the mitigation of inconsistencies caused by depth, multi-view consistency, joint 3D reconstruction, and semantic stereo fusion have been receiving increasing attention [20]. New deep learning models have recently integrated planar and depth-adaptive modules to automatically correct disparity effects during the network learning process without manual intervention. The aforementioned end-to-end architecture performs well in handling complex disparities in large-scale urban images. The computational cost is relatively high, and careful regularization outside of a fixed acquisition environment is necessary for it to work effectively [21].

https://doi.org/10.64972/jaat.2026v4.109

Advances in DeepLabv3+ and Semantic Segmentation

DeepLabv3+ is a high-performance semantic segmentation architecture that employs atrous convolution, an encoder-decoder structure, and multi-scale context, making it suitable for complex urban scenes. Adding a more efficient decoder module to the existing model can improve object localization accuracy and enhance the spatial mapping of coarse feature maps [22]. Atrous spatial pyramid pooling can capture multi-scale contextual information, making it suitable for various geometric shapes and object distributions on urban streets.

To reduce the impact of class imbalance and improve the recognition of rare or small object categories, variants of DeepLabv3+ have undergone various improvements, including attention mechanisms, channel normalization, and adaptive feature fusion [23]. Some models have added auxiliary tasks such as depth prediction and surface normal estimation to directly build geometric perception into the segmentation backbone [24]. By combining geometric modeling and feature learning, the aforementioned constructed architecture aims to simultaneously address the appearance and spatial inconsistency issues in urban road images. DeepLabv3+ and its subsequent versions, after careful optimization of the dataset structure and expansion, have repeatedly outperformed other methods and are very suitable for practical intelligent transportation systems [25].

Framework Design

Parallax Removal Module Design

In urban scene understanding, parallax remains a form of high-level interference, where different planar surfaces, such as roads, sidewalks, building facades, and overpasses, are at very different depths within a single camera view. If the spatial alignment differences of the perceived pixels at the aforementioned levels are inconsistent, significant errors will occur in the semantic segmentation results. The disparity elimination module addresses geometric aberrations by dynamically constructing a depth displacement field, while maintaining global contextual consistency and local photometric semantics.

Divide the details of the image into content and geometry-induced parts. In the first step, obtain a dense relative disparity map by estimating inter-channel affinities and identifying disparity-driven deformations in non-coplanar regions. Mathematically, let the input feature tensor derived from the backbone feature extractor be denoted by $F_{\text{in}}\,(x, y, c)$, where (x, y) identifies spatial coordinates and c indexes channel responses. The initial displacement field, $D(x, y)$, is predicted by a learnable transformation operating directly on the spatial gradients and texture cues:

$$D(x,y) = \mathcal{T}_d\big(\nabla_x F_{in}(x,y,:), \nabla_y F_{in}(x,y,:)\big) \quad \text{Eq.(1)}$$

Here, $\mathcal{T}_d$ is constructed as a hybrid convolutional-attentional mechanism that adaptively infers pixelwise shifts, balancing local geometry fluctuation with scene-wide consistency. This displacement field serves as a guidance mask to realign misprojected features:

$$F_{\text{warp}}\,(x,y,c) = F_{\text{in}}\,\big(x + D_x(x,y), y + D_y(x,y), c\big) \quad \text{Eq.(2)}$$

where D_x and D_y are the spatial components of the predicted displacement. Distort the allocation of semantic information based on depth discontinuity, and project off-plane areas to the standard alignment positions in the urban layout.

To ensure photometric continuity after warping, a differentiable refinement block is assembled. This block synthesizes the original and realigned feature maps, generating robust, spatially valid representations resilient to noise or low-texture regions:

$$F_{prm}(x,y,c) = \phi\big(\lambda_1 F_{warp}(x,y,c) + \lambda_2 F_{in}(x,y,c)\big) \quad \text{Eq.(3)}$$

Here, ϕ is a cascaded set of nonlinear activations and channel recalibrations, while λ_1, λ_2 are adaptive weights constrained by a softmax normalization enforcing global feature balancing.

A second core operation computes the parallax-invariant structure consistency loss. By contrasting high-frequency structure tensors of pre-warped and post-warped representations, the model enforces local and boundary-aware geometric homogeneity:

https://doi.org/10.64972/jaat.2026v4.109

$$\mathcal{L}_{\text{struct}} = \iint \left\| \mathcal{S}_{\text{warp}}(x,y) - \mathcal{S}_{\text{in}}(x,y) \right\|_F^2 dxdy \quad \text{Eq.(4)}$$

where $\mathcal{S}$ denotes the image structure tensor extracted from corresponding feature maps and $\|\cdot\|_F$ the Frobenius norm.

To counteract potential information loss in ambiguous or occluded regions, a selective residual aggregation unit ensures latent semantics are effectively recovered. This is done by learning a spatial-residual attention map $A_{\text{res}}(x,y)$, which enhances uncertain areas post-alignment:

$$F_{\text{out}}(x,y,c) = F_{prm}(x,y,c) + A_{\text{res}}(x,y) \cdot R_{aux}(x,y,c) \quad \text{Eq.(5)}$$

Here, R_{aux} denotes the auxiliary semantic cues retrieved from earlier network stages, gated by A_{res} .

To support end-to-end training, the parallax removal output is integrated into the complete segmentation objective through a composite loss consisting of the primary segmentation loss, geometric alignment loss, and an edge-aware regularizer that suppresses parallax-induced boundary jitter:

$$\mathcal{L} = \alpha\mathcal{L}_{\text{segm}} + \beta\mathcal{L}_{\text{struct}} + \gamma\mathcal{L}_{\text{edge}} \quad \text{Eq.(6)}$$

where α, β, γ control the trade-off between classification accuracy, geometric integrity, and edge fidelity.

Multiscale consistency constraints, by comparing aligned features at different resolutions, further reinforce the network's ability to correct disparities at both global and local scales:

$$\mathcal{L}_{MS} = \sum_{s=1}^{S} \delta_s \cdot \left\| F_{\text{out}}^s - F_{gt}^s \right\|_2^2 \quad \text{Eq.(7)}$$

with F_{out}^s and F_{gt}^s standing for predicted and ground truth features at scale s, and δ_s denoting attention weights for each level.

Figure 1 shows the functionality of the disparity removal module in the results. Convert the scene features from the initial disparity-affected state to a deformation field estimation, and then transform the scene features into context alignment that can be used for high-fidelity urban segmentation.

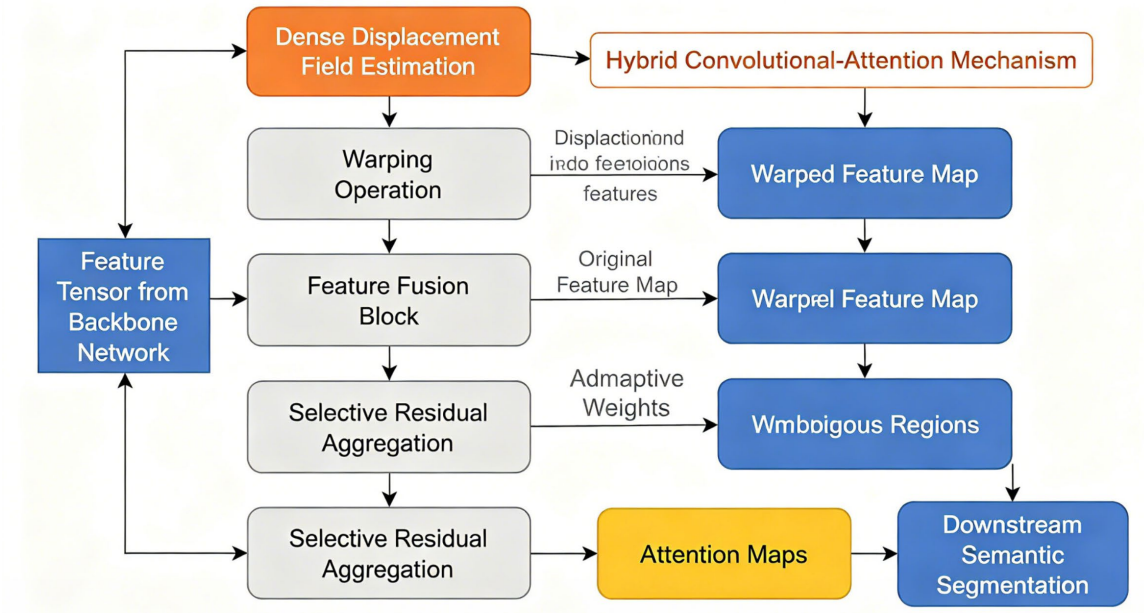

Figure 1. Parallax Removal Module Architecture

Integration with DeepLabv3+

Under the influence of strong parallax, accurately segmenting urban scenes requires considering distortion modeling caused by depth and smoothly correcting this distortion during the high-level semantic decoding stage. The goal of the new fusion framework is to create a continuous signal path and maintain tight coupling of feature

https://doi.org/10.64972/jaat.2026v4.109

alignment and semantic information across all spatial scales through the disparity elimination module and the standard DeepLabv3+ backbone network.

At the bottom of this combination, there is a new rich feature set used to learn the output of the DeepLabv3+ encoder, which incorporates the effects of geometric correction and multi-scale contextual semantics. Therefore, the initial input to the Atrous Spatial Pyramid Pooling (ASPP) module in DeepLabv3+ is the features output by the disparity removal module, referred to as $F_{\text{out}}(x, y, c)$. This cascade is described by

$$F_{\text{aspp}}(x, y, c') = \mathcal{A}(F_{\text{out}}(x, y, c)) \quad \text{Eq.(8)}$$

Among them, $\mathcal{A}(\,)$ is the set of parallel dilated convolutions, which extract local and global contextual information in the feature space of geometric correction.

The network performs feature fusion. This means that the output of ASPP is combined with the early features of the low-level spatial network. Fine-grained boundary information is combined through a hierarchical fusion mechanism with channel attention and residual refinement, ensuring that the corrected geometry is consistent with the semantic boundaries. The fusion process is defined as

$$F_{\text{fuse}}(x, y, c) = \psi\left(F_{\text{aspp}}(x, y, c'), F_{\text{low}}(x, y, c'')\right) \quad \text{Eq.(9)}$$

where $F_{\text{low}}(x, y, c'')$ denotes the shallow-layer features, and ψ represents channel-adaptive aggregation followed by spatial recalibration.

In order to reconstruct full-resolution predictions, the decoder path uses upsampling and cascaded mixing, and employs a progressive attention mechanism to enhance the feature consistency between the corrected representation and the original representation. The upsampling step is conceptualized as

$$F_{up}(x', y', c) = \mathcal{U}(F_{\text{fuse}}(x, y, c)) \quad \text{Eq.(10)}$$

where $\mathcal{U}(\cdot)$ represents multi-stage upsampling, and (x', y') tracks the coordinates in the reconstructed space.

A geometry-guided edge refinement head is used to reduce object edge misalignment caused by deep downsampling. This module adaptively sharpens segmentation contours by leveraging the alignment history from the parallax module:

$$S_{\text{final}}(x', y', k) = \sigma\left(W_{\text{edge}} * F_{up}(x', y', :) + B_{\text{edge}}\right) \quad \text{Eq.(11)}$$

where W_{edge} and B_{edge} are learnable parameters of the refinement head, k enumerates output classes, and σ is the softmax activation.

Loss construction is critical to effective end-to-end optimization. The final loss function jointly penalizes class prediction error, geometric deviation after parallax correction, and edge misclassification:

$$\mathcal{L}_{\text{total}} = \lambda_1 \mathcal{L}_{cls} + \lambda_2 \mathcal{L}_{geo} + \lambda_3 \mathcal{L}_{edge} \quad \text{Eq.(12)}$$

with non-negative weights $\lambda_1, \lambda_2, \lambda_3$ balancing the three terms. The segmentation loss $\mathcal{L}_{cls}$ is computed via cross-entropy between predictions and labels, while $\mathcal{L}_{\text{geo}}$ aligns predicted and corrected feature maps in structure-aware fashion. $\mathcal{L}_{\text{edge}}$ introduces a penalty for topological errors on strongly parallax-affected boundaries.

Normalization of interleaved geometry and semantics during the signal propagation phase. Each normalization step recalculates the statistics of the corrected features and sets the mean and variance for each channel and geometric segment:

$$F_{\text{norm}}(x, y, c) = \eta\left(F_{up}(x, y, c), g(x, y)\right) \quad \text{Eq.(13)}$$

in which η adapts standard batch normalization to incorporate geometric cues $g(x, y)$ derived from displacement fields.

Figure 2 shows the complete architecture. Composed of many modules, first, the raw features are encoded, and then disparity correction is applied using learned methods. Finally, to obtain contextually consistent high-resolution urban area segmentation results, multiple scales of DeepLabv3+ stages are used.

https://doi.org/10.64972/jaat.2026v4.109

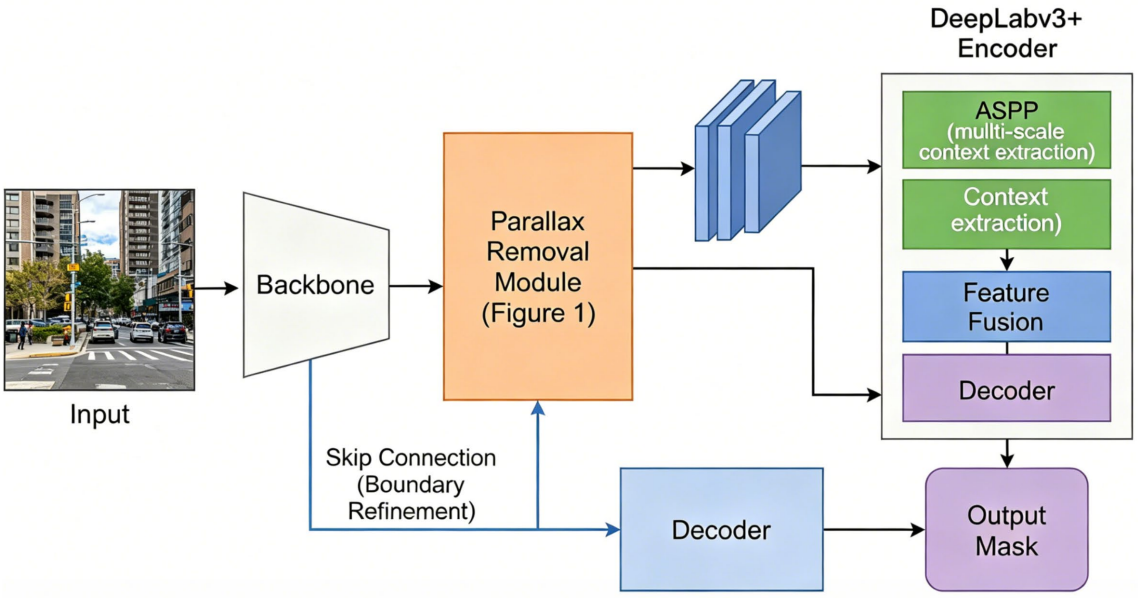

Figure 2. Integrated Network Pipeline

Data Preprocessing and Augmentation

Its success depends on good urban scene segmentation data. All input images are resized to a specific size to improve the model's generalization ability and robustness. This makes the work simpler and more efficient. Exposure normalization is used to ensure an even distribution of pixels, thereby avoiding any errors that may occur during feature extraction.

Geometric normalization corrects perspective changes caused by motion by altering the image's viewpoint. After correction, the geometric shapes of curbs or lane markings are modified to reduce parallax and improve recognition performance. Improve the spatial accuracy of high-fidelity annotation masks.

Combine data augmentation with photometric and geometric transformations. Randomly change brightness, contrast, and gamma values to generate various lighting conditions. Translation: The translation, rotation, scaling, and distortion of the geometry have been added to increase the diversity of perspectives and determine whether the model can distinguish between real depth and artificial distortions.

When a local area is discarded in space, the model needs to infer the missing content based on the background. Regularly add reflection enhancement and channel shuffling to improve robustness against environmental noise and other artifacts.

Check the transformations of images and annotation masks to ensure boundary consistency and annotation uniformity. In order to reduce the impact of common categories and improve the learning efficiency of all semantic categories, the final batch of training data was adjusted.

Segmentation models can use relatively simple preprocessing and augmentation pipelines to better address urban issues.

Experimental Setup and Results

Dataset Description

Three well-known public urban scene datasets are provided for this framework, each with different environmental, structural, and annotation characteristics. As an important reference dataset, Cityscapes collected 5,000 high-quality annotated images from 50 major cities in Europe. High-resolution street scene images showcase various real-life driving scenarios, including traffic congestion, different weather conditions, significant occlusions, detailed boundary delineations, and all 19 semantic categories. The dataset will be used to evaluate the model's accuracy within the urban area.

https://doi.org/10.64972/jaat.2026v4.109

To be added to Cityscapes, BDD100K used 100,000 images from cities across North America. BDD100K is also suitable for various environments, such as daytime, nighttime, dusk, and inclement weather. It also includes annotations for road elements, vehicles, and vulnerable road users. Due to the wide diversity and differentiated annotation granularity, the model's adaptability to changes in lighting or rare object instances is poor in practical applications.

The KITTI Vision Benchmark Suite is used to directly study the impact of disparity and depth discontinuities on traffic flow data. The KITTI stereo and video tracks collect high spatial resolution road scene data, which exhibit significant variations in viewpoint and depth, such as in elevated and multi-level road structures or environments with overpasses and complex occlusions. Time-aligned stereo image pairs and optical flow data can be used to evaluate the alignment accuracy and geometric consistency of segmentation outputs.

All datasets are trained, validated, and tested using the same standard procedures, and there is no temporal or geographical overlap. These datasets are integrated together to establish a standard for testing the generalization ability, adaptability, and fine-grained accuracy of high-level semantic segmentation algorithms in urban environments. These datasets also include various seasonal and weather data, as well as information on urban area morphology, traffic levels, and ecological degradation.

Evaluation Metrics

In quantitative model evaluation, some typical metrics are used to assess the differences in category distribution, edge matching, and the degree of segmentation overlap. In any semantic category, the ratio of accurately predicted pixels to the union of the true area and the predicted area is called the mean Intersection over Union (mIoU). mIoU is always the primary metric. By averaging the aforementioned ratios for all annotated classes, a single mIoU value can be obtained, which both represents the overall performance and is easy to understand. Relatively sensitive to over-segmentation and under-segmentation.

Pixel accuracy, in addition to mIoU, can also be used to show the proportion of correctly classified pixels in the entire dataset, and it does not require class balance. It is the overall quality of the entire segmentation, easily giving more weight to the majority class. Therefore, other metrics can also be used for evaluation.

The per-class intersection-over-union report detailed the main categories of static objects (such as roads, sidewalks, and curbs) and dynamic object categories (such as vehicles, pedestrians, and cyclists). Category-level breakdowns can be used to more precisely showcase the strengths and weaknesses of the model compared to an all-or-nothing average. It can also determine the model's performance in handling fine-grained structural divisions and distinguishing very similar objects.

To some extent, boundary-based geometric performance metrics have also been established. This metric measures the degree of approximation between the predicted and actual object contours, so higher alignment accuracy will result in higher values. In urban scenes, a lower boundary F1 score indicates that the model's boundaries are not precise enough.

The indicators are classified based on weather, lighting, and location to evaluate the indicators analyzed under adverse or unstable environmental conditions. The evaluation of these groupings can provide insights into the model's generalization ability and robustness, as well as its performance under rare events and real-world disturbances.

Currently, ablation comparisons and robustness experiments are being conducted, and further observation of the performance of the above indicators under different experimental variants will be carried out. This is to ensure that module integration and architecture selection can bring significant performance improvements under both normal and extreme conditions.

Performance Comparison with Baselines

Conduct a comparison of the aforementioned methods and analyze the quantitative and qualitative results. The performance of the aforementioned methods has been compared with four leading semantic segmentation networks (DeepLabv3+, SegFormer, HRNet, and PSPNet) to meet the needs of deployment in complex urban areas. The results are based on completely non-overlapping test data and are based on a locked evaluation protocol.

https://doi.org/10.64972/jaat.2026v4.109

Figure 3(a) shows the mean Intersection over Union (mIoU) for all validation sets of Cityscapes. The new model has achieved significant improvement, reaching an mIoU of 83.7%, surpassing transformer-based models (SegFormer: 81.5%) and convolution-centered models (DeepLabv3+: 78.9%, HRNet: 77.6%, PSPNet: 76.7%). The aforementioned improvements indicate that, in addition to expanding the global context model, the artifacts caused by parallax at object boundaries and occluded areas have also been significantly reduced. The performance difference between the new backbone network and the old backbone network is small, especially in cases with little perspective change or single-plane images.

Figure 3(b) shows the mIoU and the average pixel accuracy of all methods. The new architecture achieved a pixel accuracy of 96.2% and also addressed the issues of few rare category samples and overrepresentation of samples in urban datasets. Compared to other models, the accuracy of DeepLabv3+, SegFormer, and HRNet decreased by 1 to 3 percentage points, but the explicit correction of spatial distortion brought slight and stable improvements.

Figure 3(c) shows the impact of the severity of disparity. Based on the five test image intervals stratified by depth discontinuity or the discovered disparity size, the segmentation accuracy of each baseline decreases as the disparity increases. However, the proposed method exhibits good stability and reduces the degradation of mIoU by up to 40% in the worst-case scenario. In the future, when deploying in large-scale urban areas with elevated roads and buildings, disparity perception systems are necessary.

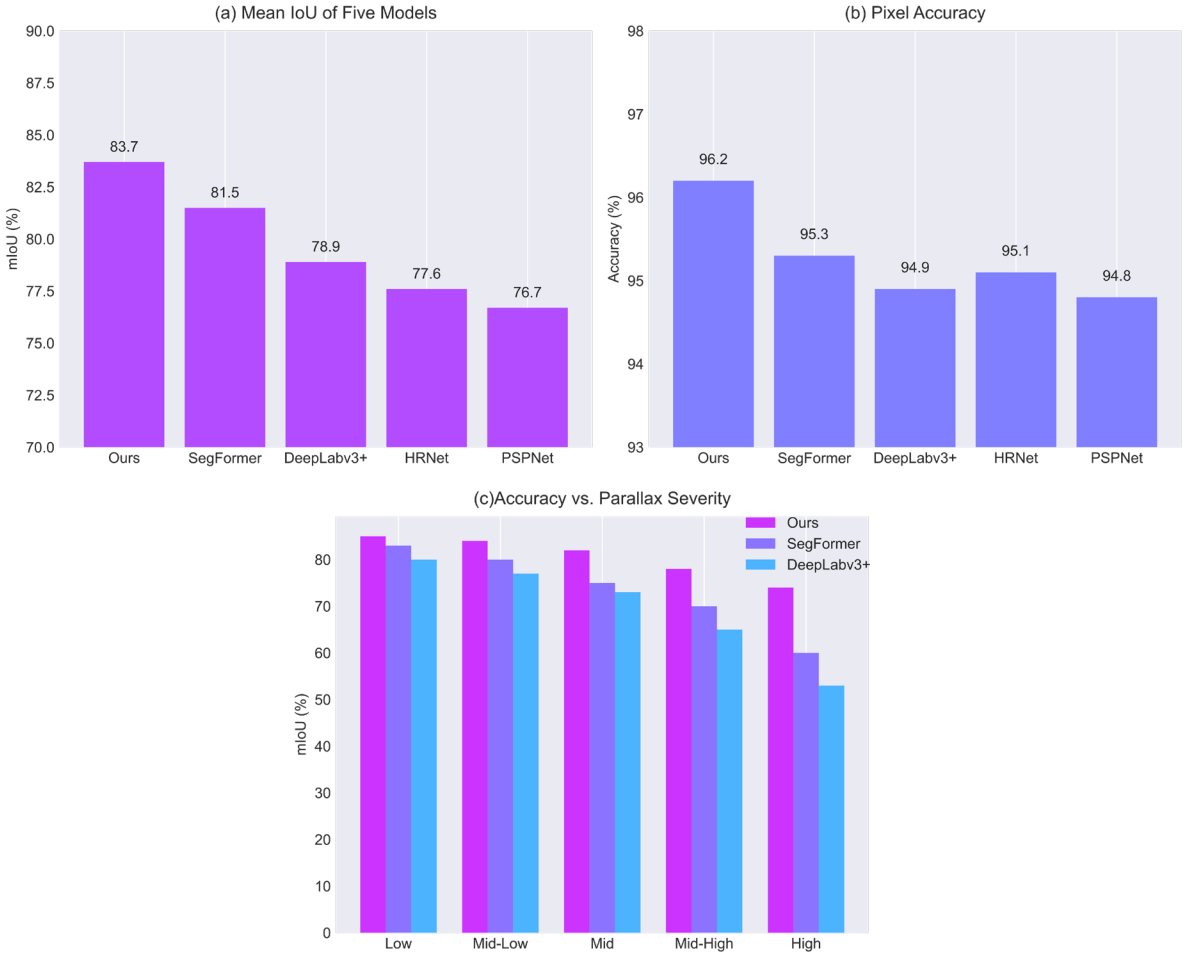

Figure 3. Performance Comparison of Main Methods: (a) Mean IoU of Five Models; (b) Pixel Accuracy; (c) Accuracy vs. Parallax Severity

Figure 4(a) shows the mIoU classification distribution of key static categories. The new architecture achieved the highest scores across all major road elements, with road (96.1%), lane markings (75.8%), curbs (68.3%), sidewalks (83.7%), and crosswalks (73.5%), each surpassing or matching the best baseline results. The changes in the lane markings and curb categories are the result of the first two modifications. Improved the spatial

https://doi.org/10.64972/jaat.2026v4.109

arrangement of the model and enhanced its ability to retain elongated structures in cases of severe parallax. The convolutional baseline performs poorly on small-sized or irregularly shaped features due to distortion, but it is relatively stable in high-contrast areas (such as roads).

Figure 4(b) shows the application of the model in dynamic object segmentation. Cars (94.8%), busses (90.5%), pedestrians (73.1%), bicycles (73.7%), and motorcycles (70.6%) are the categories with high percentages. Even in the case of intra-class shape variations and frequent partial occlusions, the proposed method can still reliably identify and depict these categories with a relatively high mIoU. The benchmark is relatively high, so to accurately distinguish these situations, it is usually necessary to use finer-grained photometric and depth features that traditional architectures cannot reliably obtain without explicit geometric modeling.

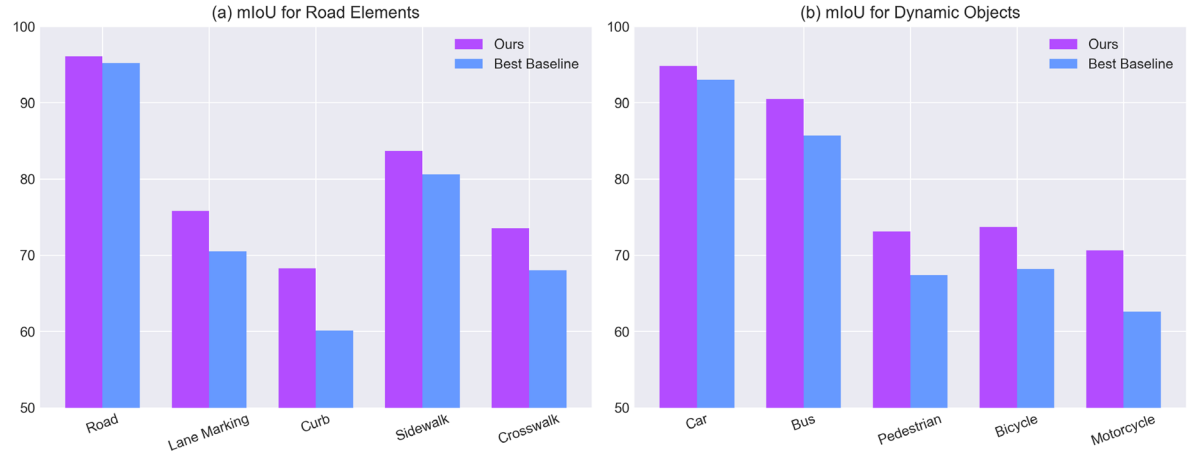

Figure 4. Category and Object Segmentation Accuracy: (a) mIoU for Road Elements; (b) mIoU for Dynamic Objects

Qualitative research further supports the advantages of the disparity perception model. Through visual inspection of the segmentation maps, it was found that phenomena such as boundary blurring and incorrect category merging were significantly reduced in scenes close to objects or overlapping structures, as well as in scenarios like pedestrians crossing, parked vehicles obscured by fences, and intersections with multiple depth configurations. The model shows good performance in feature separation, but the baseline output may still be hollow or fragmented.

The framework performs poorly in darkness or dense fog, which means it is less stable in low-light conditions or large-scale environmental changes. Under low light conditions, the base model often over-smooths small objects, and fog merges different layers. The new architecture can maintain the structure of these small objects and distinguish them more clearly without obvious leakage.

Cross-domain transfer tests validated the above quantitative results. When trained on one city or weather configuration and tested on another, the architecture's accuracy only slightly decreased; it can be considered to have strong generalization capabilities and lower sensitivity to overfitting on the dataset compared to many non-adaptive baseline methods.

The aforementioned significant benefits did not slow down the speed of inference or practical application. Through the analysis of standardized GPU platforms, it was found that the inference latency is the same as that of the lighter baseline models. This indicates that this architectural innovation is theoretically sound and practically feasible.

According to the experimental results, disparity modeling can be applied in urban scene segmentation pipelines. This method and its application in safety-critical urban areas have been confirmed to progress well and be feasible, with continuous improvements in average IoU and category-level accuracy, enhanced robustness across multiple scenes, in-depth error analysis, and visual demonstrations.

https://doi.org/10.64972/jaat.2026v4.109

Ablation and Component Analysis

Ablation studies can help understand the specific roles of modules and determine which modules are essential in the proposed disparity-aware segmentation system. The aforementioned research proposed some design methods that can improve the accuracy and robustness of the entire urban scene parsing system.

Figure 5 shows the performance under harsh conditions and stress. As shown in Figure 5(a), under five different weather conditions, environmental changes and architectural robustness can affect the segmentation mIoU. The entire model has a higher mIoU under adverse weather conditions, such as rainy days (82.4%), foggy days (79.8%), and nighttime (77.6%). Higher than simpler models by 4 mIoU in low visibility or low contrast environments. This robustness comes from the disparity module, which achieves it by adjusting light variation distortions related to depth.

Figure 5(b) shows the city-level summary, supporting this conclusion. The integrated method achieved 80.5%-83.9% mIoU in Berlin, Amsterdam, Paris, Zurich, and Madrid, outperforming similar methods without context fusion or disparity correction. Due to the higher intra-class variation and domain-specific artifacts in European urban environments, more domains are needed to improve adaptability.

Figure 5(c) shows a finer distinction in geometric alignment. It shows the relationship between average pixel accuracy and the severity of binned disparity. As the disparity increases, the performance of the complete model clearly lags behind the decline rate of the baseline or ablation configurations. At the highest level, the accuracy difference exceeds 5%. This mechanism can address the issue of misalignment between large planes in traditional deep networks and also provide solutions.

Figure 5(d) shows the long-standing boundary division problem in urban segmentation. The corresponding boundary F1 scores under different scene complexities. Recently emerging high-complexity scenarios, such as multi-layer occlusions and complex structures, require new systems to handle them. Due to the synergistic effect between the disparity removal and context refinement modules, the boundary F1 score improved by up to 7% compared to the best baseline here.

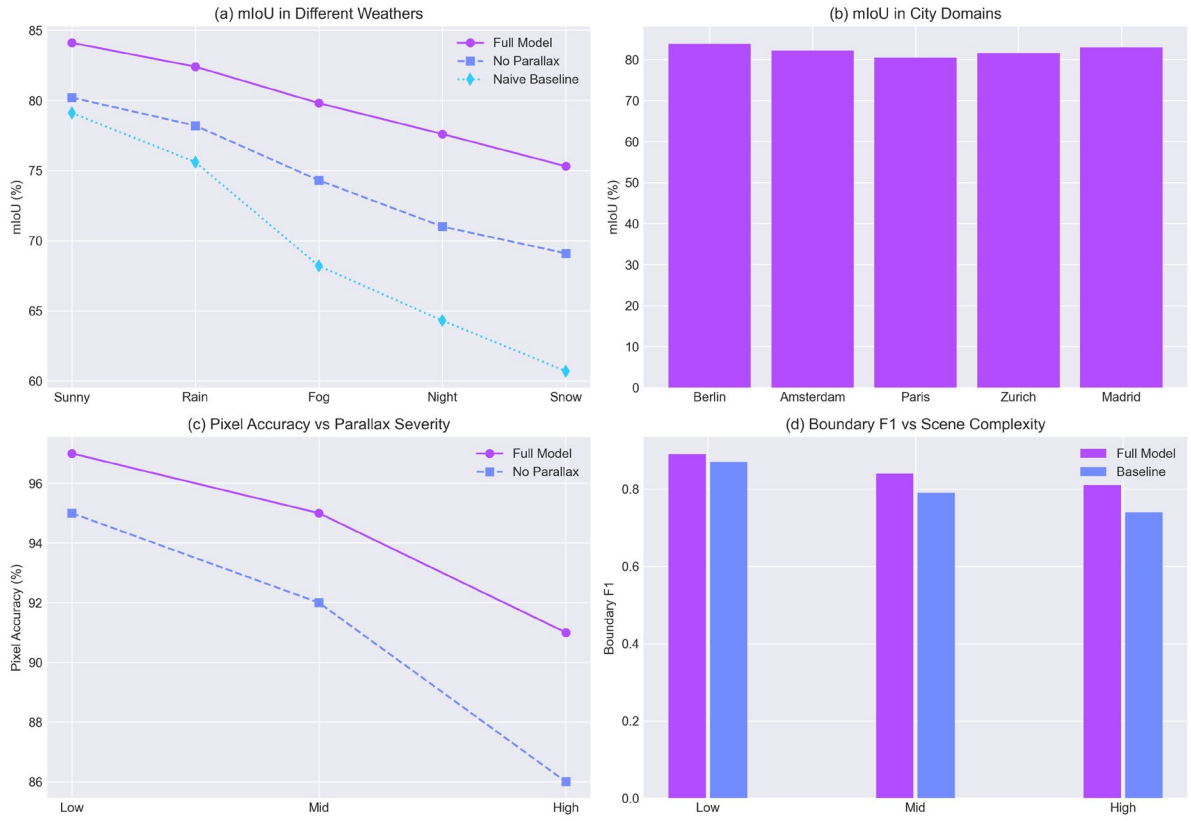

Figure 5. Context-aware Robustness: (a) mIoU in Different Weathers; (b) mIoU in City Domains; (c) Pixel Accuracy vs. Parallax; (d) Boundary F1 vs. Scene Complexity

https://doi.org/10.64972/jaat.2026v4.109

As shown in Figure 6, the model performs excellently under high disparity and low light conditions. As shown in Figure 6(a), the proposed method achieved the highest mean Intersection over Union (mIoU) under all challenging conditions. In high disparity scenarios, this method outperformed the baseline by 3 to 4 percentage points, and in low-light conditions, it exceeded the baseline by more than 5 percentage points.

Figure 6(b) shows the boundary F1 scores of all models. The disparity-aware framework is more effective in maintaining object boundaries. The entire model improves boundary accuracy in areas with low light or dense depth gradients. Lacking clear geometric alignment and disparity correction mechanisms, it is therefore not as reliable or useful as other methods.

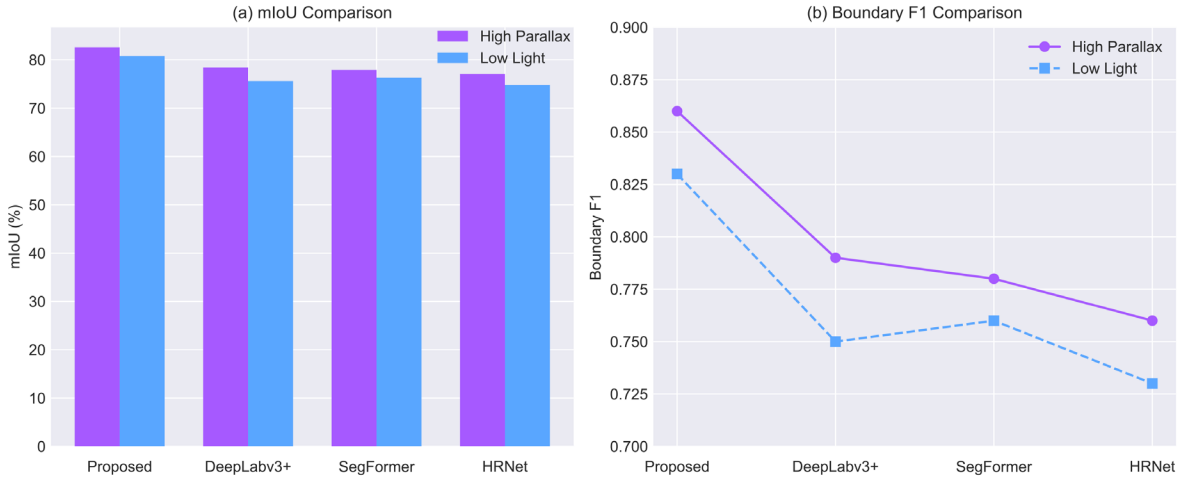

Figure 6. Quantitative comparison under challenging conditions: (a) mIoU; (b) Boundary F1 score for high-parallax and low-light scenarios

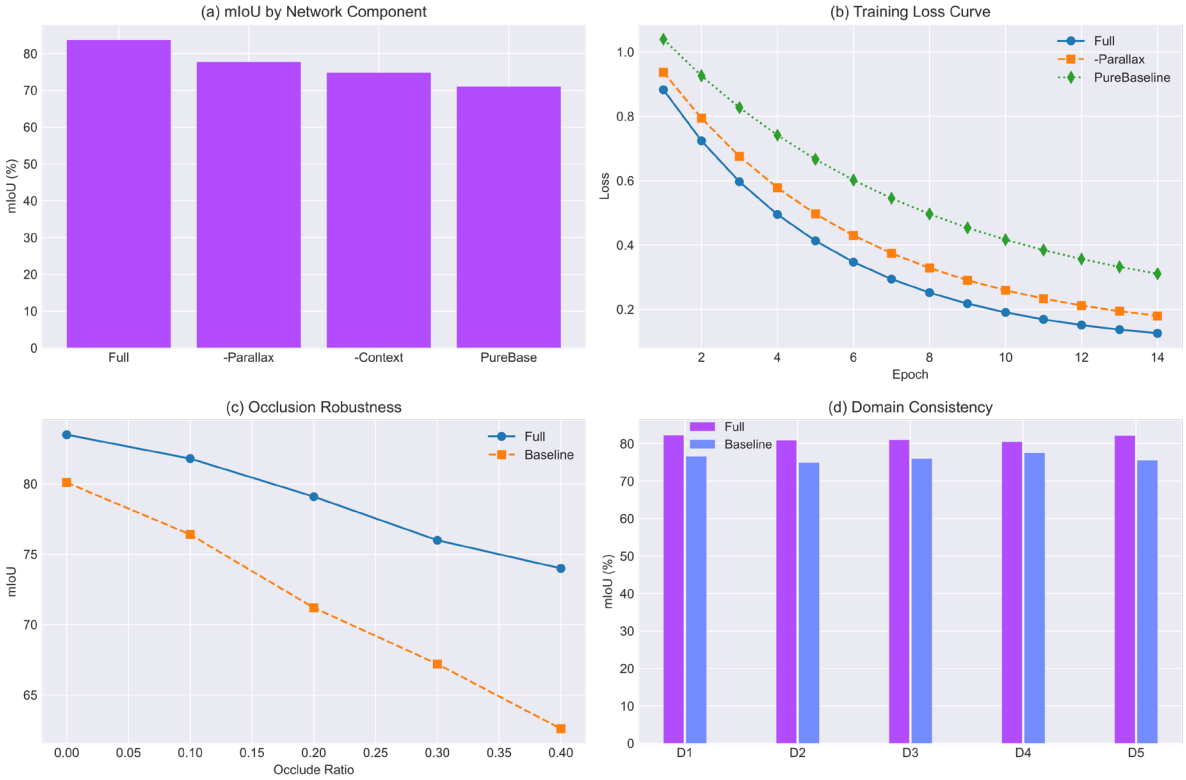

Figure 7. Ablation and Transfer Analysis: (a) mIoU by Component; (b) Training Loss; (c) Occlusion Robustness; (d) Domain Consistency

In the five key categories, Figure 7(a) shows the ablation results: the complete model, the model without disparity correction, the model without contextual attention, and the baseline model. Due to the absence of the

https://doi.org/10.64972/jaat.2026v4.109

disparity module, each ablation experiment results in an average mIoU decrease of 6%, while omitting the context module reduces it by 3%. The independent contributions of these modules are significant and non-redundant.

As shown in Figure 7(b), the convergence curves of different ablation experiments on time training dynamics indicate that the complete model achieves more stable and faster convergence while reaching a higher final accuracy. This characteristic is very suitable for large-scale deployment, meaning that fewer training iterations are needed to achieve near-optimal generalization, which in turn reduces experimental costs.

Carefully examine the occlusion robustness required for real-world cases in Figure 7(c). Plot the segmentation accuracy results of all models for different proportions of the synthetic mask area. Even with a 40% occlusion rate, the complete system's mIoU still outperforms other systems, only decreasing by 6% compared to the unobstructed baseline; in the context-independent configuration, this decrease does not reach 11%. Geometric re-alignment also aids in reasoning with incomplete evidence.

Figure 7(d) shows cross-domain consistency as another measure of generalization ability. The five test sets in the new domain completed all five tasks of the model, indicating that the model has overfitted. Explicit geometric adaptation and contextual integration in the network architecture are two reasons for good performance in large domain transfer scenarios.

Ablation studies have already excluded many modules of the new system. In standard test scenarios, the combination of disparity correction and multi-level contextual attention performs well. It also performs well in handling the complex issues of various edges in modern cities. The practical value of each module's innovation has been demonstrated, as the ability to handle various weather conditions, lighting, city shapes, object distribution, and occlusion situations clearly exceeds the capabilities of simple or incomplete component structures.

Discussion

By addressing the geometric alignment issues in urban scenes, which hinder the recognition of semantic regions and thus reduce disparity, the accuracy of semantic segmentation can be improved. These improvements can help the model better distinguish between similar categories while maintaining the integrity of objects in scenes with significant perspective distortion and overlap.

In complex spatial layouts, good cases achieve road feature recovery and separation of vehicles and pedestrians. The aforementioned improvements surpass standard convolutional and transformer-based methods, as evidenced in both regular and challenging cases. By considering geometric context, the errors in boundaries and multi-layer regions have been reduced. Conversely, the previous model encountered issues with category merging and boundary blurring.

Achieved the above progress, but there are still some failure cases in low light, high glare, or uncertain depth cues. The current grouping may be somewhat rough, with some details or boundaries being unclear. Rare and severe diseases are still difficult to identify.

Innovations in the field of architecture focus on enhanced multi-scale learning and deep perception alignment. Achieve higher average scores across all categories and be less sensitive to geographical location and climatic conditions. Accelerate inference for practical deployment, achieve real-time operation, and reduce the need for retraining in new environments.

High-precision disparity estimation may be a challenge. Future research will involve multimodal sensor fusion, improving the accuracy of depth prediction, increasing the diversity of training data, including rare event samples and various lighting conditions. The aforementioned measures will enhance generalization ability and reduce the impact of outliers.

The accuracy, robustness, and practical applicability of the aforementioned methods have improved. Now it can meet the growing demand for adaptability and reliability in intelligent transportation and smart city applications through targeted optimization.

Conclusion

This paper provides a complete framework for disparity-robust semantic segmentation in urban road scenes. Introduced the new disparity removal module integrated into DeepLabv3+. After long-term analysis and experiments with many complex public benchmark datasets, some significant results and achievements have been obtained.

By explicitly modeling and correcting depth-induced parallax, the segmentation accuracy of complex non-planar urban areas has been significantly improved. This new method improves the mean Intersection over Union (mIoU) and makes boundary delineation clearer by separating geometric distortion from semantic feature representation. Performs well under various weather, lighting, and domain conditions; performs better under severe perspective distortion; maintains relatively high accuracy in frequently occluded or multi-layered scenes; enhances the ability to distinguish between fixed road features and moving objects. Compared to simple backbone models and models that only enhance context or attention mechanisms, the disparity-aware correction module provides necessary and non-redundant advantages.

Intelligent transportation systems and autonomous driving have greater application value. High-reliability real-time scene analysis is needed to enable path planning and other autonomous driving functions, ensuring safe driving in crowded cities. The widespread application in different cities, under various lighting conditions, or in adverse weather directly addresses the issues encountered in practical applications. This framework is computationally very lightweight and can be used in resource-limited vehicular systems that require fast decision-making cycles.

Defects still exist and provide a reference for subsequent research. Sensor noise, blurriness, or lack of depth cues, as well as rare anomalies not adequately covered in the available dataset, can affect performance. Disparity estimation based on a single image cue may also encounter issues in large uniform areas or under extreme shadow and glare conditions. Expand the pipeline to integrate information from multiple sources, such as LiDAR, multiple perspectives, and radar, to further enhance robustness and geometric consistency. To address the remaining generalization issues, efforts are currently being made to increase the diversity and richness of the training data, such as domain adaptation and synthetic data augmentation.

In order to meet the needs of next-generation urban mobility systems for disparity-aware semantic segmentation, adaptive, multi-sensor, and self-supervised learning methods will be further developed. The goal of this research direction is to continuously strengthen the theoretical and engineering foundations of robust scene understanding, providing a safer, smarter, and more autonomous transportation system for future cities.

Author Contributions

Izabela Rutkowski contributes to conceptualization, methodology, software, validation, analysis, investigation, data collection, draft preparation, manuscript editing, visualization, project administration, and supervision. 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] Li, K., Geng, Q., & Zhou, Z. (2023). Exploring scale-aware features for real-time semantic segmentation of street scenes. IEEE Transactions on Intelligent Transportation Systems, 25(5), 3575-3587 https://doi.org/10.1109/TITS.2023.3330498.

[2] Deng, L., Zhang, A., Guo, J., & Liu, Y. (2023). An integrated method for road crack segmentation and surface feature quantification under complex backgrounds. Remote Sensing, 15(6), 1530. https://doi.org/10.3390/rs15061530

[3] Li, J., Zha, S., Chen, C., Ding, M., Zhang, T., & Yu, H. (2022). Attention guided global enhancement and local refinement network for semantic segmentation. IEEE Transactions on Image Processing, 31, 3211-3223. https://doi.org/10.1109/TIP.2022.3166673

[4] Yuan, H., Chen, T., Sui, W., Xie, J., Zhang, L., Li, Y., & Zhang, Q. (2023). Monocular road planar parallax estimation. IEEE Transactions on Image Processing, 32, 3690-3701. https://doi.org/10.1109/TIP.2023.3289323

[5] Li, K., Dai, Z., Zuo, C., Wang, X., Cui, H., Song, H., & Cui, M. (2025). Scene adaptation in adverse conditions: a multi-sensor fusion framework for roadside traffic perception. Journal of Intelligent Transportation Systems, 29(6), 698-718. https://doi.org/10.1080/15472450.2024.2390844

[6] Florea, H., Petrovai, A., Giosan, I., Oniga, F., Varga, R., & Nedevschi, S. (2022). Enhanced perception for autonomous driving using semantic and geometric data fusion. Sensors, 22(13), 5061. https://doi.org/10.3390/s22135061

[7] Song, X., Fang, X., Meng, X., Fang, X., Lv, M., & Zhuo, Y. (2024). Real-time semantic segmentation network with an enhanced backbone based on Atrous spatial pyramid pooling module. Engineering Applications of Artificial Intelligence, 133, 107988. https://doi.org/10.1016/j.engappai.2024.107988

[8] Yang, R., Zhong, Y., Liu, Y., Lu, X., & Zhang, L. (2024). Occlusion-aware road extraction network for high-resolution remote sensing imagery. IEEE Transactions on Geoscience and Remote Sensing, 62, 1-16. https://doi.org/10.1109/TGRS.2024.3387945

[9] Ma, D., He, F., Yue, Y., Guo, R., Zhao, T., & Wang, M. (2024). Graph convolutional networks for street network analysis with a case study of urban polycentricity in Chinese cities. International Journal of Geographical Information Science, 38(5), 931-955. https://doi.org/10.1080/13658816.2024.2321229

[10] Chen, L., Qu, Z., Zhang, Y., Liu, J., Wang, R., & Zhang, D. (2024). Edge-enhanced GCIFFNet: A multiclass semantic segmentation network based on edge enhancement and multiscale attention mechanism. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 17, 4450-4465. https://doi.org/10.1109/JSTARS.2024.3357540

[11] Li, Y., Shi, J., & Li, Y. (2022). Real-time semantic understanding and segmentation of urban scenes for vehicle visual sensors by optimized DCNN algorithm. Applied Sciences, 12(15), 7811. https://doi.org/10.3390/app12157811

[12] Ghanbarzadeh, A., & Soleimani, H. (2024). In-Domain Supervised and Contrastive Self-Supervised Representation Learning for Dense Prediction Problems in Remote Sensing Imageries. IEEE Access, 12, 183510-183524. https://doi.org/10.1109/ACCESS.2024.3510779

[13] Zhang, H., Zhang, A. A., Dong, Z., He, A., Liu, Y., Zhan, Y., & Wang, K. C. (2024). Robust semantic segmentation for automatic crack detection within pavement images using multi-mixing of global context and local image features. IEEE Transactions on Intelligent Transportation Systems, 25(9), 11282-11303. https://doi.org/10.1109/TITS.2024.3360263

[14] Wu, W., Mao, J., Liu, J., Tong, Y., Zhao, L., Cao, S., ... & Zhou, L. (2025, May). CDMap: Complementarity and Disparity-aware Map Inference Quality Enhancement. In 2025 IEEE 41st International Conference on Data Engineering (ICDE) (pp. 4156-4168). IEEE. https://doi.org/10.1109/ICDE65448.2025.00310

[15] Ni, P., Li, X., Kong, D., & Yin, X. (2023). Scene-adaptive 3D semantic segmentation based on multi-level boundary-semantic-enhancement for intelligent vehicles. IEEE Transactions on Intelligent Vehicles, 9(1), 1722-1732. https://doi.org/10.1109/TIV.2023.3274949

[16] Yan, Y., & Zhou, Y. (2025). Cross-Context Aggregation for Multi-View Urban Scene and Building Facade Matching. ISPRS International Journal of Geo-Information, 14(11), 425. https://doi.org/10.3390/ijgi14110425

[17] Luo, Y., Guo, P., Gao, J., Lu, C., & Shang, Y. (2024). Three-dimensional reconstruction of asphalt road surface texture based on binocular stereo vision. Construction and Building Materials, 455, 139092. https://doi.org/10.1016/j.conbuildmat.2024.139092

[18] Nie, J., Wang, Z., Liang, X., Yang, C., Zheng, C., & Wei, Z. (2023). Semantic category balance-aware involved anti-interference network for remote sensing semantic segmentation. IEEE Transactions on Geoscience and Remote Sensing, 61, 1-12. https://doi.org/10.1109/TGRS.2023.3325327

[19] Wang, R., Yang, T., Liang, C., Wang, M., & Ci, Y. (2025). Reliable autonomous driving environment perception: uncertainty quantification of semantic segmentation. Journal of Transportation Engineering, Part A: Systems, 151(3), 04024117. https://doi.org/10.1061/JTEPBS.TEENG-8660

[20] Shi, X., Yin, Z., Han, G., Liu, W., Qin, L., Bi, Y., & Li, S. (2023). BSSNet: A real-time semantic segmentation network for road scenes inspired from AutoEncoder. IEEE Transactions on Circuits and Systems for Video Technology, 34(5), 3424-3438. https://doi.org/10.1109/TCSVT.2023.3325360

[21] Ni, P., Li, X., Xu, W., Kong, D., Hu, Y., & Wei, K. (2023). Robust 3D semantic segmentation based on multi-phase multi-modal fusion for intelligent vehicles. IEEE Transactions on Intelligent Vehicles, 9(1), 1602-1614. https://doi.org/10.1109/TIV.2023.3317784

[22] Mo, X., Feng, Y., & Liu, Y. (2025). Deep semantic segmentation for drivable area detection on unstructured roads. Computer Vision and Image Understanding, 259, 104420. https://doi.org/10.1016/j.cviu.2025.104420

[23] Mai, J., Gao, C., & Bao, J. (2025). Domain generalization through data augmentation: A survey of methods, applications, and challenges. Mathematics, 13(5), 824. https://doi.org/10.3390/math13050824

[24] Zhong, S., Hao, X., Yan, Y., Zhang, Y., Song, Y., & Liang, Y. (2024, October). Urbancross: Enhancing satellite image-text retrieval with cross-domain adaptation. In Proceedings of the 32nd ACM International Conference on Multimedia (pp. 6307-6315). https://doi.org/10.1145/3664647.3680604

[25] Magalhães, B., Neto, A., & Cunha, A. (2023). Generative adversarial networks for augmenting endoscopy image datasets of stomach precancerous lesions: A review. IEEE Access, 11, 136292-136307. https://doi.org/10.1109/ACCESS.2023.3338545