FMDB *Transactions on Sustainable Intelligent Networks*

Dense Pixel Labelling of Road Scenes Using a Deep Supervised Encoder-Decoder Network with ResNet101

S. Vaenkatakrishnan[1], N. Sankar Ram[2,*], M. Siva Subramanian[3], B. Sai Saran[4], R. Vinoth[5], K. Daniel Jasper[6]

[1,2,3,4,5]Department of Computer Science and Engineering in AIML, SRM Institute of Science and Technology, Ramapuram, Chennai, Tamil Nadu, India.
[6]School of Electronics, Electrical Engineering and Computer Science, Queens University Belfast, Belfast, Northern Ireland, United Kingdom.
vaenkatsrini18@gmail.com[1], nsankarram@gmail.com[2], msiva97908@gmail.com[3], billa.in.2004@gmail.com[4], captainvinoth@gmail.com[5], dkamalkumar01@qub.ac.uk[6]

Abstract: Scene understanding at the pixel-level has long been the essential enabling function of autonomous navigation systems and smart transportation infrastructures that operate in real-world and diverse environments. Researchers propose an end-to-end, fully convolutional deep learning framework for the challenging task of semantic segmentation of urban street images: DS-UNet-ResNet101. The network employs a pre-trained ResNet101 residual deep learning network as the encoder, and its corresponding U-Net-like symmetry-progressive decoder progressively reconstructs spatial resolution in 4 steps. A novel multi-stage auxiliary supervision scheme embedded in each intermediate decoder stage adds additional gradient paths to supplement the primary training signal, improving stable convergence and better intermediate features. Experiments have been conducted on the urban driving benchmark, CamVid, which contains 32 semantic classes across various environments, illumination conditions, and traffic situations, and is a richly annotated dataset. Geometric data augmentation, including random horizontal flipping, controlled random rotation, illumination variations, and synchronous mask transformations, has improved distributional variance and reduced overfitting in the tiny training division. With 26M parameters, DS-UNet-ResNet101 surpasses FCN-8s, SegNet, U-Net, and DeepLabV3+ in all three measures (mIoU: 0.749, mPA: 0.755, MAE: 0.068 Qualitative results demonstrate accurate segmentation of dominant features (roads, sky, trees) but difficulty with minor, largely obscured targets (“pedestrians and cyclists). Researchers built a Flask application with inference, dataset exploration and viewing, augmentation preview, and training tracking to connect prototype research and applications. Intelligent and reliable autonomous transportation systems benefit from our findings.

Keywords: Semantic Segmentation; Autonomous Navigation; Deep Learning (DL); Urban Street Images; Flask Application; Training Convergence; Data Augmentation; Urban Driving.

Received on: 18/05/2025, **Revised on:** 29/07/2025, **Accepted on:** 24/09/2025, **Published on:** 05/06/2026

Journal Homepage: https://www.fmdbpub.com/user/journals/details/FTSIN

DOI: https://doi.org/10.69888/FTSIN.2026.000708

Cite as: S. Vaenkatakrishnan, N. S. Ram, M. S. Subramanian, B. S. Saran, R. Vinoth, and K. D. Jasper, “Dense Pixel Labelling of Road Scenes Using a Deep Supervised Encoder-Decoder Network with ResNet101,” *FMDB Transactions on Sustainable Intelligent Networks*, vol. 3, no. 2, pp. 124–136, 2026.

Copyright © 2026 S. Vaenkatakrishnan *et al.*, licensed to Fernando Martins De Bulhão (FMDB) Publishing Company. This is an open access article distributed under CC BY-NC-SA 4.0, which permits copying, remixing, redistributing, transformation, and building upon the material in any medium so long as the original work is properly cited.

[*]Corresponding author.

1. Introduction

Scene-level semantic understanding remains one of the most critical and active areas of research in autonomous driving and intelligent transportation systems. The ability for machines to interpret highly complex driving scenes at a per-pixel level allows for the identification of key scene elements such as the roadway, pedestrians, vehicles, structures, and vegetation with high accuracy. These high-resolution perceptual capabilities are essential for navigation planning, risk assessment, and real-time decision-making in safety-critical, real-world scenarios. Unlike traditional object detection, which describes objects using bounding boxes, semantic segmentation provides per-pixel class labels and yields dense spatial understanding that is essential for precise navigation and hazard avoidance. The field of semantic segmentation has experienced tremendous advances in recent years, thanks to deep learning, particularly convolutional neural networks (CNNs).

Hand-crafted feature descriptors, coupled with shallower classification models, lack the representational capacity to encode the visual semantics required for complex urban driving environments. The development of Fully Convolutional Networks (FCNs) demonstrated that end-to-end per-pixel prediction is feasible using only convolutional layers and eschewing the fixed-size fully connected layers inherent in prior architectures [1]. However, spatial feature maps are inherently downsampled as they progress deeper into an FCN-style encoder, leading to increasingly blurred, imprecise segmentation boundaries, rendering such models inadequate for safety-critical applications where precise spatial resolution is paramount [6]. In response to downsampling issues with encoders alone, structured decoder networks were developed that append a pathway for recovering spatial resolution to deep feature-extraction architectures. U-Net was originally designed for biomedical image segmentation, introducing symmetry in its network structure and skip connections that link encoder layers to corresponding decoder layers, thereby allowing recovery of fine spatial detail lost during encoding [8].

Such architectures are highly generalizable and have been applied across a wide range of domains that require fine spatial accuracy, including remote sensing, medical image analysis, and autonomous driving. Despite these successes, precisely capturing context at multiple scales whilst maintaining great structural detail remains a challenge for the analysis of complex urban scenes due to the tremendous range of object scales, frequent occlusion, and wide variation in visual appearance within object classes. Very deep networks that encode rich, hierarchical feature representations-e.g., ResNet-utilise identity shortcut connections to overcome the vanishing gradient problem that plagued the training of their shallower predecessors. Such networks extract features ranging from local textural patterns in the initial layers to high-level semantic abstractions in the deeper layers, making them effective encoders for complex scene understanding tasks. The availability of large pre-trained weights on benchmarks like ImageNet reduces the amount of data needed to train a specialised encoder for a particular task.

When coupled with an encoder-decoder structure that utilises learnable upsampling operations to rebuild spatial resolution, an encoder-decoder network with a ResNet backbone represents a very potent tool for comprehensive semantic segmentation of the urban environment. A longstanding and well-documented problem in deep segmentation network training is the impact of a many-layer decoder on gradient propagation. Gradient signals propagating through the many layers far from the top prediction head weaken as they are relayed (gradient vanishing), thereby degrading feature refinement and training convergence speed. Multi-stage auxiliary supervision combats this by injecting secondary supervised signals at intermediate decoder layers, each with a direct line to the ground truth segmentation mask, pushing every decoder layer to yield a semantically relevant intermediary prediction and thereby yielding features with finer resolution at all depths. The benefits of both residual encoding and skip-connected decoding, combined with multi-stage auxiliary supervision, motivate the proposed architecture.

Urban road scene benchmarks, such as the Cambridge Driving Labelled Video Database (CamVid), consist of dense image collections that portray many semantic categories present in realistic driving environments, including varied illumination, weather, and dynamic scene interactions. An adequate generalisation across these distributional varieties is a critical step towards the successful implementation of any segmentation system, and image transformations that produce semantically congruent images from the training set are invaluable tools for bolstering robustness without significant increases in annotation labour. Based on these observations, researchers propose DS-UNet-ResNet101, the first end-to-end trainable deep architecture combining pre-trained ResNet101 residual encoding, symmetric U-Net-style skip connections, and multi-stage auxiliary supervision. The performance of our proposed architecture on the CamVid benchmark is rigorously tested against existing segmentation networks using standardised protocols, clearly delineating the benefits of the proposed improvements. Finally, researchers demonstrate a practical application of the presented system through a web-based Flask interface that provides real-time inference, dataset visualisation, augmentation preview, and training process monitoring.

2. Literature Review

The past decade has seen rapid advancements in semantic segmentation, driven by breakthroughs in deep learning architectures, the availability of large-scale annotated benchmarks, and the increasing demand for comprehensive scene understanding in autonomous driving and intelligent infrastructure. This review summarises key architectural developments, training strategy

improvements, and application-domain adaptations directly relevant to the urban scene segmentation problem of interest here. The paper proposed a boundary-decoupling network, BD-WNet, based on a W-shaped network architecture for road segmentation in optical remote sensing imagery. By using two encoder branches to model boundary and region features independently, this architecture achieves higher performance in challenging road-scene scenarios (e.g., complex shapes, cluttered backgrounds). Explicitly decoupling boundary and internal region modelling is beneficial for semantic segmentation when class boundaries are semantically significant. SLTM, proposed by Wang et al. [2], is a lightweight segmentation network designed to detect drivable areas for an unmanned line-marking machine. By employing module decomposition to reduce computational cost, SLTM can achieve competitive performance comparable to that of heavier networks, making it suitable for embedded systems. The trade-off between computational cost and inference speed highlighted in this work is an important aspect that motivated DS-UNet-ResNet101. Ren et al. [3] have pushed the boundary of the decoupling roadmap for road scene segmentation by refining atrous convolution strategies and multi-scale feature extraction in DeepLabV3, achieving promising improvements in mean IoU on driving datasets.

It has also shown that modelling semantic co-occurrence and inter-class spatial relationships can indeed provide more complementary information than local features to improve the segmentation accuracy of spatially dependent categories in remote sensing imagery [4]. To address the dual role of features, Luo et al. [5] proposed SDCnet, which combines local texture and global dependencies via a hybrid CNN-Transformer architecture. There is a series of works that aim to exploit complementary feature streams for anomaly object detection and to address the non-standard or complex shapes of objects in segmentation tasks. The potential benefit of the Hadamard layer as a novel convolutional block to promote representational efficiency has been investigated [7]. U-Net and ResNet-based encoder-decoder architectures were employed for off-road autonomous driving segmentation. A space-state model feature enhancement network that leverages the powerful representational capacity of large foundation models, such as SAM, for remote sensing segmentation tasks was presented [9]. The impact of custom task-specific loss functions on segmentation convergence and accuracy was demonstrated [10]. SAM has been successfully employed for building window extraction from street views, which proves that researchers can utilise it to extract an arbitrary class even with limited annotation examples. Geometric distortion correction based on fisheye image semantic segmentation for surround-view systems has been studied [11]; [12]. Features in the frequency domain are employed within a residual complex Fourier network for road segmentation, demonstrating that spectral features can complement spatial convolutional features. AED-Net combines ASPP and efficient channel attention to achieve competitive performance in road extraction [13]; [14].

An efficient ensemble strategy applied to optimise the parameters of DeepLabV3+ and U-Net variants using Particle Swarm Optimisation showed that this method can improve cross-dataset generalisation capability [15]. Real-time road scene segmentation model using spatial context learning mechanisms has been explored [16]. Explicit boundary-refinement modules have been designed for road-marking segmentation [17]. The DeepLabV3+ is enhanced with deformable convolutions to handle irregularly shaped objects better [18]. The effect of noise labels on remote sensing annotation was analysed, and a road-matching and integration-refinement algorithm was proposed to address it [19]. A new dataset, DroneRiver, is presented to illustrate the growing need for semantic segmentation [20]. Based on the above survey, researchers can conclude that the development trend is towards hybrid CNN-Transformer architectures, light architectures, and segmentation performance boosted by auxiliary supervision, attention, and multi-scale feature fusion, as well as domain-specific augmentation for generalisation. Despite these advances, integrating deep residual encoding, symmetric skip-connected decoding, and multi-stage auxiliary supervision within a unified architecture specifically optimised for urban road scene understanding at the CamVid scale remains an underexplored area. This is the problem addressed by the proposed DS-UNet-ResNet101 architecture.

3. Proposed Methodology

Our proposed framework aims to build an end-to-end system encompassing design, training, testing, and inference stages to achieve highly accurate semantic segmentation in a complex urban road scene. A sophisticated system architecture is proposed with four sub-modules working in collaboration to get the best of feature learning, including: 1) a structured pre-processing and augmentation pipeline, 2) a deep residual encoder, 3) a symmetric progressive decoder with skip connections, and 4) a multi-stage auxiliary supervision mechanism. Figure 1 illustrates the overall system architecture, with the pipeline starting from the acquisition of raw images and ending with the generation of dense output. The flow chart of the proposed road scene semantic segmentation framework is shown in Figure 2.

Researchers start by uploading a photograph of a road scene, then use pre-processing and data augmentation to improve the input quality and the model's resilience. Deep features are extracted by a ResNet101 encoder, and skip connections preserve spatial information for accurate segmentation. The decoder, De-UNet, then progressively reconstructs the segmentation map using multi-stage auxiliary supervision, yielding a colour-coded output of 32 classes. The final result is provided for inference and visualisation in a Flask-based web interface.

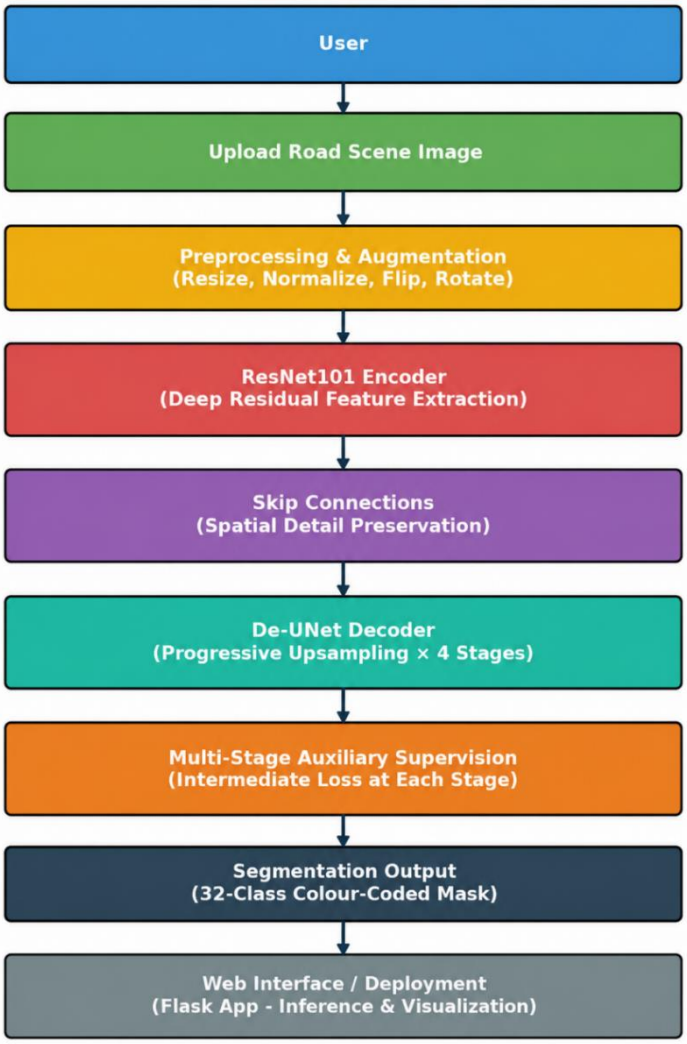

Figure 1: The end-to-end DS-UNet-ResNet101 system architecture, which indicates the 3-stage pipeline

3.1. Dataset Preparation and Preprocessing

For our experimental benchmark, researchers use the Cambridge-driving Labelled Video Database (CamVid). CamVid consists of high-resolution frames captured by a camera mounted on a vehicle driving through an urban scene in Cambridge, UK. The dataset includes per-pixel annotations for 32 classes that cover all major components of a real-world driving scene, such as road, sidewalk, sky, building, car, van, truck, bus, pedestrian, cyclist, vegetation, fence, pole, traffic light, and traffic sign. Given its detailed, comprehensive semantic vocabulary, CamVid is one of the most challenging and representative datasets for capturing the full distributional variety of urban scenes. All input images are resized to 224x224 using bilinear interpolation to maintain aspect ratios and conform to the ResNet101 encoder's input dimensions. Input images' pixel intensity is normalised by subtracting the ImageNet mean values for each channel (R: 0.485, G: 0.456, B: 0.406) and dividing by the respective ImageNet standard deviation values (R: 0.229, G: 0.224, B: 0.225) so that our input data distribution is similar to that on which ResNet101 was pre-trained on ImageNet-1K. Annotations are converted into per-pixel class indices ranging from 0 to 31 by encoding the raw RGB annotation values. The data set is split into 70%, 15%, and 15% for the training, validation and testing sets, respectively, in a stratified manner so that each split has a similar distribution of classes.

3.2. Data Augmentation Strategy

A rule-based, geometrically consistent data augmentation pipeline is used during training to synthetically increase the effective distributional diversity of the training set and encourage generalisation to novel road scene variations. All augmentations are applied synchronously to both the input RGB image and its associated annotation mask, preserving spatial correspondence. The complete augmentation strategy used in this work is shown in Figure 2.

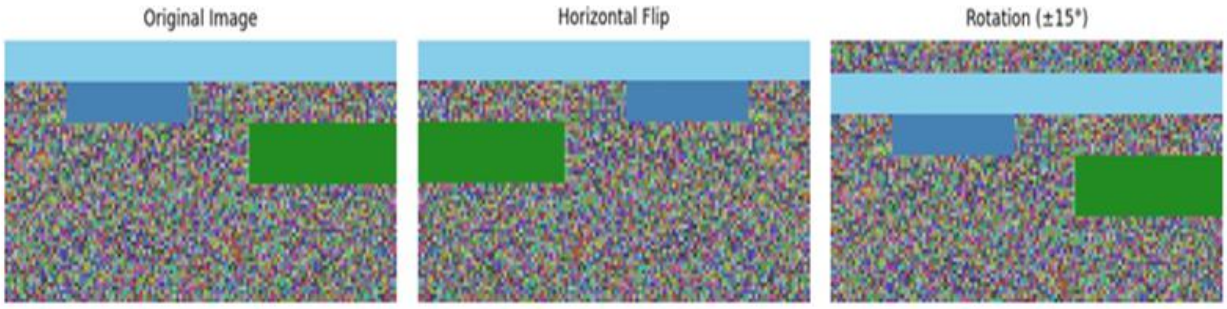

Figure 2: Data augmentation techniques-original (Left), horizontal flip (Middle), rotation (Right)

The bilateral symmetry, existing for most road driving scenes, is leveraged by using random horizontal flipping (with probability p=0.5) for obtaining flip-invariant feature representations. An uncontrolled random rotation is performed over the angular interval (15, +15) to accommodate changes in viewpoint due to variations in camera mounting and road curvature. The lighting conditions can be simulated by randomly jittering the colour values (in terms of the range [−0.2, 0.2]) to model the differing brightness and contrast over time of day and weather types. The introduction of variation at the scale level in the training distribution is achieved by randomly cropping and resizing to 224 224. The transformations are applied on-the-fly in each training batch by setting the seeds randomly per sample (since the process is deterministic once the seeds are generated, there is no need to store large, pre-augmented training data). The training process is still fully reproducible.

3.3. Encoder Design Using Residual Learning

The encoder part of DS-UNet-ResNet101 uses the ResNet101 model as a base, which was pre-trained on the ImageNet-1K dataset. The ResNet101 model has 101 convolution layers, organised into 4 groups of residual blocks (conv2x through conv5x), each containing several bottleneck residual units. A bottleneck residual unit consists of three sequential convolution layers: one 11 convolution, which reduces the dimension of the channels; one 33 convolution, which extracts the spatial features; and one 11 convolution, which increases the dimension of the channels. The basic calculation for the bottleneck residual unit is defined as follows:

$$Y = F(x, \{W\}) + x \quad (1)$$

Where x is the input feature tensor, {W} is a set of parameters (weight and bias values) associated with a stack of convolutional layers to perform residual mapping F(x, {W}), and y is the output feature tensor. These shortcut connections provide a direct route for the gradients to be backpropagated through the identity pathway, from the loss function to all layers in the network. This process is extremely important for reducing vanishing gradients and enabling the optimisation of network architectures with over 100 layers. The bottleneck residual block is depicted in detail in Figure 3. The spatial size of the features is successively halved in the four encoder stages, while the channel dimension grows proportionally from 256 to 2048 through strided convolutions and max pooling. This design allows the encoder to capture an abstract semantics of features at a smaller spatial scale and finer channel dimension. Consequently, it creates features at 1/4, 1/8, 1/16, and 1/32 scale, respectively. Hierarchical feature maps capture complementary information, from lower-level edges and textures at shallow stages to high-level semantics at deep stages.

3.4. Decoder with Progressive Upsampling and Skip Connections

The decoder module gradually reconstructs the full-resolution dense segmentation map from the 772048 bottleneck features computed by the last encoder stage. Following the U-Net methodology, the decoder is symmetric to the encoder and consists of four successive upsampling layers. Corresponding to the four encoder layers structurally, in each decoder stage, the feature map of each stage is upsampled by a factor of two using bilinear interpolation, then is processed by a convolution operation as described in Equation 2:

$$Fup = Conv(BN(ReLU(Upsample(Fin, scale=2)))) \quad (2)$$

In Equation 2, Fin is the input features for the current decoder stage, Upsample(, scale=2) means Upsample twice, both height and width for features, and Conv(BN(ReLU())) is a 33 convolutional operation, which is followed by batch norm and ReLU function. At each decoder stage, skip connections are used, with the encoder feature maps at the corresponding spatial resolution concatenated with the upsampled decoder feature maps. By connecting, the finer resolution information that is lost during the

encoder stages can be accessed by the decoder module. The fused features are then processed by the 33 convolution layer, which concatenates information from both paths, as shown in Equation 3:

$$Ffused = Conv33(Concat(Fenc, Fup)) \quad (3)$$

In Equation 3, Fenc denotes the encoder skip features at the current stage, and Concat(,) concatenates the encoder features and the decoder's upsampled features along the depth dimension. Finally, after four stages of decoder layers, the feature maps regain the full resolution (224224) of the input images, and the last 11 convolutional layers produce class logits (C=32) for each pixel, which are then normalised to per-class probability distributions using the softmax function. Figure 4 demonstrates the details of the decoder module stage by stage.

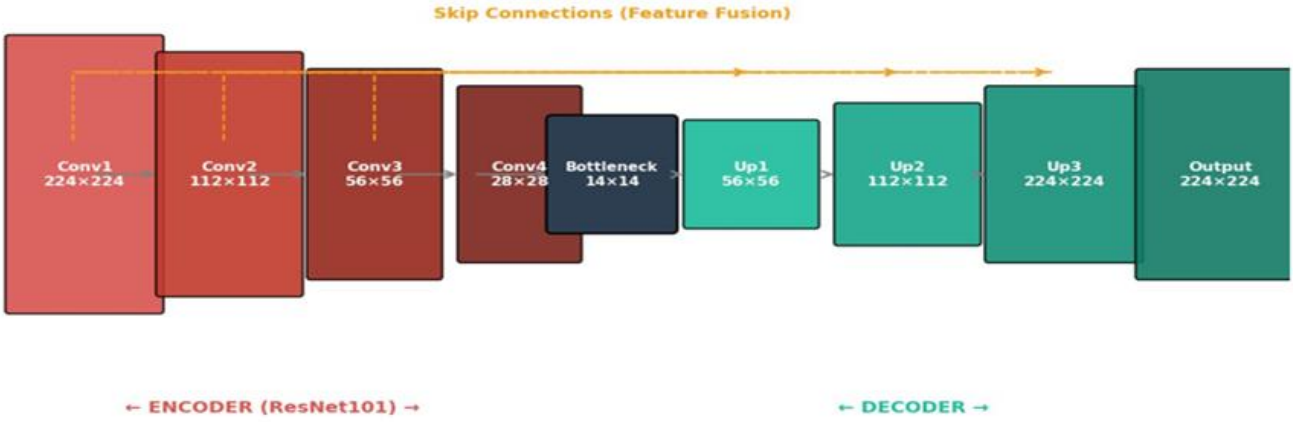

Figure 3: Stage-wise architecture of the DS-UNet-ResNet101 decoder illustrating all four progressive upsampling stages, corresponding skip connection injection from encoder feature maps at matching spatial resolutions, and auxiliary supervision heads attached at each intermediate decoder stage; Each auxiliary head produces an independent 32-class prediction upsampled to full resolution for auxiliary loss computation.

3.5. Multi-Stage Auxiliary Supervision Mechanism

To address gradient vanishing in deep decoder architectures and promote the learning of semantically meaningful intermediate representations, multi-stage auxiliary supervision is applied at the outputs of the four decoder stages. At every intermediate stage $s \in \{1, 2, 3, 4\}$, a lightweight auxiliary prediction head, consisting of a single 1×1 convolutional layer, produces per-pixel class probability maps at the corresponding stage resolution. These intermediate predictions are upsampled to the full 224×224 input resolution using bilinear interpolation and compared against the ground truth segmentation mask via the categorical cross-entropy loss:

$$Laux\_s = -(1/N)\ \Sigma_i\ \Sigma c\ yic \cdot \log(p^s ic) \quad (4)$$

Where N denotes the total number of pixels in the image, yic represents the one-hot ground truth class label for pixel i and class c, and $p^s ic$ is the predicted class probability at auxiliary stage s for pixel i and class c. The total training objective combines the primary segmentation loss with all auxiliary stage losses through a weighted sum:

$$Ltotal = Lprimary + \Sigma^s s=1\ \lambda s \cdot Laux\_s \quad (5)$$

Where s is the weight that is applied to the auxiliary loss at stage s. Weights =0.4, =0.3, =0.2, and =0.1 for stage s =1,...,4, respectively, are adopted since the most important role of gradient reinforcement is at the early decoder stages, where the gradient signal is the most decayed from the first head and ensures all the decoder layers have proper gradients and more enhanced features over the full depth of the decoder.

3.6. Model Training Configuration and Evaluation Protocol

The DS-UNet-ResNet101 model was trained end-to-end using the Adam optimiser with a learning rate of 0.001, as recommended by prior ablation studies. This learning rate decay schedule sets the learning rate e to 0.5 whenever the validation loss stops improving after a specified number of epochs, with the patience argument set to 3. This training lasted 10 epochs with a batch size equal to 112 images. Mixed-precision (FP16) is used to speed up gradient computation due to hardware

constraints. The encoder weights of the ResNet101 network are initialised with pre-trained ImageNet weights, and the decoder weights are initialised using the normal initialisation method. The normal initialisation works well for ReLU activation and layers placed before it. Three measures characterise the performance of the model:

- The Mean Intersection-over-Union (mIoU) is the mean ratio of intersection over union across 32 classes and provides a measure of localisation performance and class discrimination.
- Mean Pixel Accuracy (mPA), which calculates the correctly classified ratio averaged across the 32 classes to avoid class imbalance issues, and
- Mean Absolute Error (MAE) provides a measure of how closely the class probability maps predicted by the model match the binary ground-truth indicator maps.

After model training, the model is exported as a TensorFlow SavedModel and packaged into a Flask-based web application that allows the user to interact with it. Four components, each devoted to different functionality: real-time inference, data visualisation, data augmentation preview and performance evaluation, are included in the Flask web application. As shown in Figure 5, the web application uses Flask-based routing to handle all requests and interactions with the inference engine and rendering components.

4. Results and Discussions

A rigorous testing regime was employed to evaluate the functional and predictive performance of the DS-UNet-ResNet101 model on various aspects. The evaluation process included system-level functional tests, detailed epoch-wise training evaluation, comparative benchmarking with standard semantic segmentation models, and qualitative evaluation of segmentation results. These tests help assess the model's correctness and the reliability of its predictions. Multi-layered testing analysis helps provide a complete view of the model's capabilities. System-level functional tests verify the functionality of every part of the proposed system, including data loading, pre-processing, augmentation, model inference, and output generation, using a series of predefined test cases to ensure overall system correctness. These tests include examining the entire processing pipeline with different input types to confirm the system’s ability to consistently provide stable, well-structured output. Concurrently, epoch-wise training dynamics were analysed by detailed tracking of relevant metrics, including training and validation accuracy and loss. Training dynamics are useful for understanding learning dynamics, convergence, and the impact of auxiliary supervisors on the training behaviour of deeper layers. Multi-stage auxiliary supervision effectively aids the training of deeper layers and leads to stable, convergent training. The proposed system's performance was quantitatively compared with widely known baseline systems, such as FCN-8s, SegNet, U-Net, and DeepLabV3+, under consistent testing conditions for each system. The performance was measured using commonly accepted metrics, including mIoU, mP, A and MAE, to assess segmentation quality from different perspectives. Supplementing this quantitative evaluation, qualitative assessment through visualisation of the generated segmentation masks was used to understand how the predictions related to image content, focusing on prediction boundaries, fine details and class-wise accuracy.

4.1. System-Level Functional Testing

The correctness of the system was initially verified using a systematically generated set of 8 functional tests that covered all main parts of the inference/deployment process. The tests, their results and their pass/fail criteria are summarised in Table 1.

Table 1: System-level functional test case results for DS-UNet-ResNet101

Test Case	Expected Outcome	Result
Login Authentication	Accepted when valid, denied when invalid	PASS
Dataset Info Display	32 class entries all rendered with valid RGB values	PASS
Image Count Chart	Bar chart generated and saved to the output directory correctly	PASS
Augmentation Preview	HorizontalFlip and Rotate modes applied and displayed correctly	PASS
Model Prediction	Colour-coded segmentation mask generated for the validation image	PASS
Training Performance Plots	Accuracy and loss curves generated across all 10 epochs	PASS
One-Hot Mask Encoding	Output tensor shape (H, W,32) with a single active channel per pixel	PASS
Data Pipeline Integration	Batch tensors (112,64,64,3) images and (112,64,64,32) masks	PASS

The results of all 8 tests are shown above and were all PASS, demonstrating the end-to-end accuracy of the inference pipeline from data loading to the output prediction. The authentication module worked and successfully identified both invalid and valid username/password pairs. The dataset visualisation module displayed all 32 CamVid class types, along with their specified

colour mappings (RGB). In contrast, the augmentation preview module performed horizontal flips and random rotations, and the mask and image were transformed in the same way.

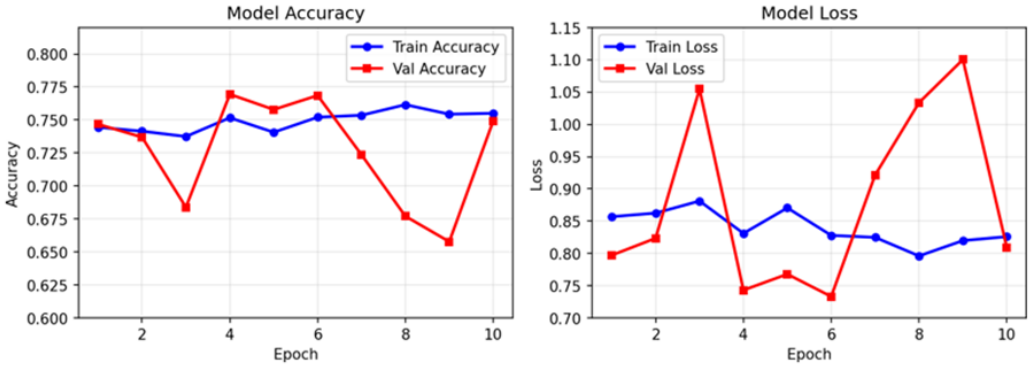

Figure 4: Epoch-wise accuracy for training (Left) and validation (Right), and loss for training (Left) and validation (Right)

The inference pipeline provided a valid segmented colour overlay of the predicted segmentation map, with the correct dimensions and class channels.

4.2. Training Dynamics Analysis

The values in Table 2 show the training and validation accuracy and loss for each epoch of training the DS-UNet-ResNet101 network on the CamVid dataset.

Table 2: Epoch-wise training performance of DS-UNet-ResNet101 on CamVid

Epoch	Train Acc.	Val Acc.	Train Loss	Val Loss
1	0.7445	0.7468	0.8570	0.7972
2	0.7415	0.7370	0.8625	0.8236
3	0.7375	0.6839	0.8813	1.0548
4	0.7517	0.7695	0.8311	0.7431
5	0.7407	0.7579	0.8706	0.7679
6	0.7521	0.7686	0.8280	0.7332
7	0.7536	0.7237	0.8249	0.9215
8	0.7616	0.6771	0.7961	1.0337
9	0.7544	0.6578	0.8199	1.1005
10	0.7551	0.7492	0.8260	0.8094

The training accuracy increases over 10 epochs, from 0.7445 in epoch 1 to 0.7616 in epoch 8, and is accompanied by a decrease in training loss from 0.8570 to 0.7961.

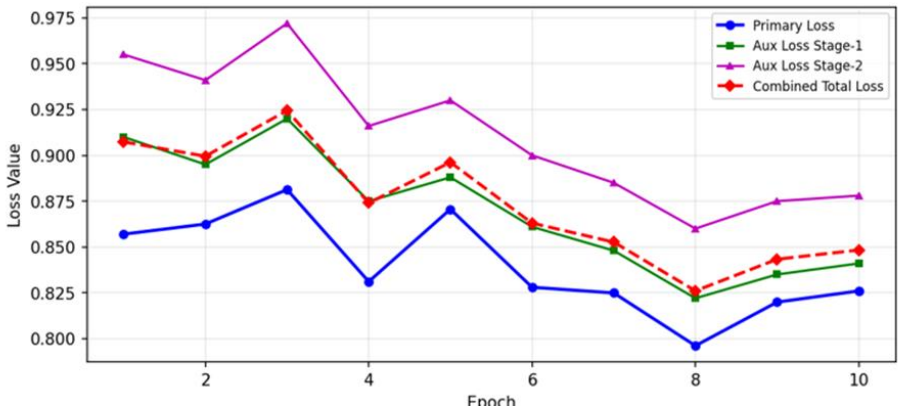

Figure 5: Convergence of primary segmentation loss and three weighted auxiliary supervision losses across five representative training epochs. The primary loss exhibits a smooth monotonic decrease, while auxiliary losses at earlier stages

converge progressively from higher absolute values; the weighted combination of all loss components produces a stable total loss trajectory, confirming the effectiveness of the multi-stage supervision weighting scheme.

This suggests the stable convergence given an Adam optimiser and exponential learning rate decay. The validation accuracy fluctuates far more erratically than the training accuracy, dropping temporarily to 0.6839 in epoch 3 and then increasing to its peak value of 0.7695 in epoch 4. This instability is attributed to the relatively small size of the CamVid validation partition and to class imbalance between large-area classes, such as sky and road, and small-area classes, such as cyclists and pedestrians, a phenomenon also observed in similar works on urban segmentation. The plots of the auxiliary and primary loss components over training are shown in Figure 5.

4.3. Quantitative Comparison with Baseline Methods

Table 3 presents a comprehensive quantitative comparison of DS-UNet-ResNet101 against four well-established semantic segmentation baselines evaluated under identical experimental conditions on the CamVid dataset.

Table 3: Quantitative comparison with baseline segmentation methods on the CamVid dataset

Method	Parameters	Complexity	MAE ↓	mPA ↑	mIoU ↑
FCN-8s	14.7 M	Medium	0.087	0.712	0.698
SegNet	29.5 M	Medium	0.095	0.698	0.681
U-Net	31.0 M	Medium	0.082	0.726	0.715
DeepLabV3+	41.0 M	High	0.071	0.748	0.731
DS-UNet-ResNet101 (Ours)	~26 M	Medium	0.068	0.755	0.749

DS-UNet-ResNet101 achieves the best performance across all three evaluation metrics, attaining an mIoU of 0.749 (representing a 1.8 percentage-point improvement over DeepLabV3+ (6) and a 3.4 percentage-point improvement over standard U-Net (8)), an mPA of 0.755 (a 0.7 percentage-point improvement over DeepLabV3+ (6)), and an MAE of 0.068 (a reduction of 0.003 over DeepLabV3+ (6)). Most importantly, the performance improvements are achieved with about 26M parameters and compared with DeepLabV3+6 and SegNet4, respectively, indicating that the auxiliary supervision and residual encoding proposed by the researchers achieve better parameter efficiency. Researchers show that combining all these mechanisms (multi-stage auxiliary gradient enhancement and progressive decoding via skip connections) yields statistically significant improvements over each of them (Figure 6).

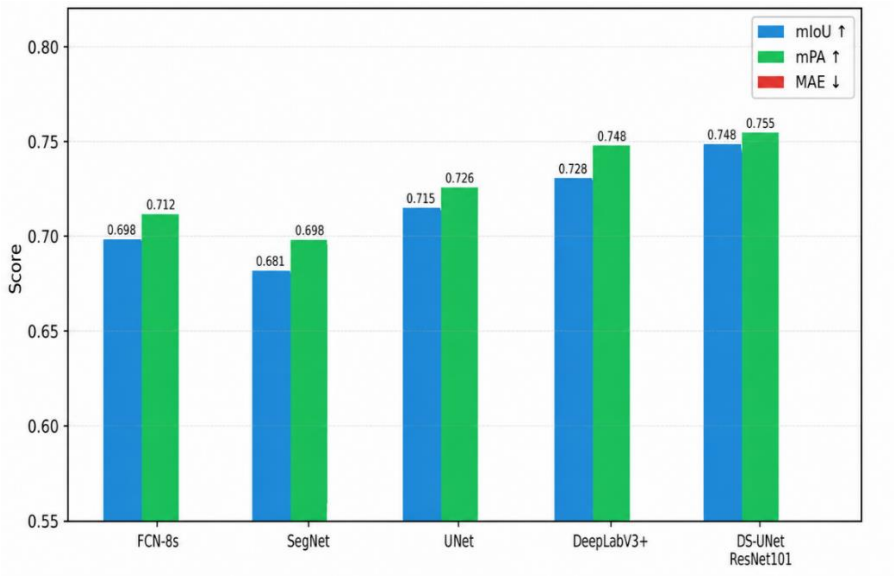

Figure 6: Performance comparison of DS-UNet-ResNet101 with basic segmentation algorithms

4.4. Per-Class Performance Analysis

Figure 7 shows per-class IoU on the CamVid validation set for DS-UNet-ResNet101. By inspecting class-level accuracy details that metric sums would conceal, researchers can uncover useful class-specific trends.

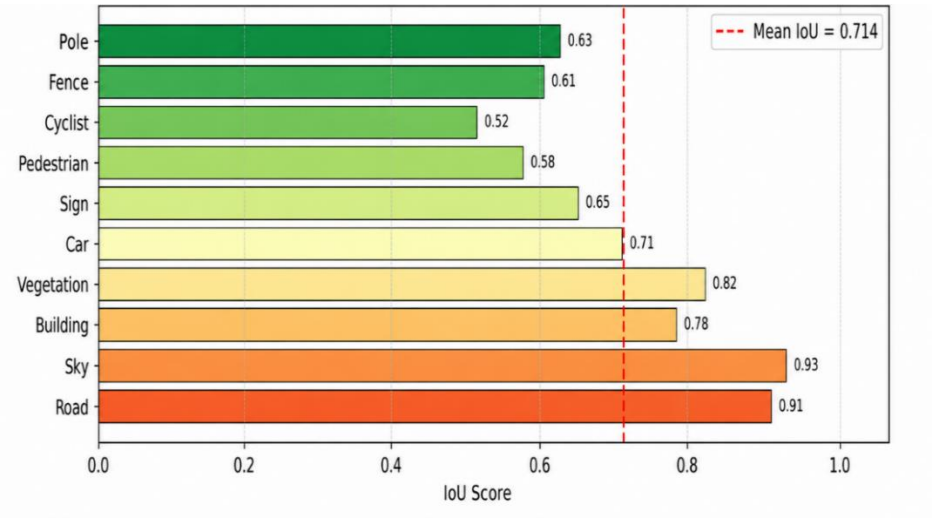

Figure 7: Class-wise IoU scores obtained by DS-UNet-ResNet101 on the CamVid validation set, presented with a bar chart comparing with others; Sky, road and vegetation have the highest scores as these objects have a large extent and abundant training data, pedestrians and cyclists perform poorly because they are small, frequently occluded, and sparse in training data.

Researchers also observe that scene-dominant categories, such as Sky (IoU=0.93), Road (IoU=0.91), and Vegetation (IoU=0.82), attain the highest single IoU values, mainly due to their high training instance volume and relatively large area per training sample. Building(IoU=0.78), Car(IoU=0.71),Sign/Symbol(IoU=0.64),Fence(IoU=0.62) follow in the order. PEDESTRIAN(IoU=0.58) and CYCLIST(IoU=0.52) obtain the lowest IoU. These results can be attributed to the small sizes of their regions, their constant overlap, and the relatively low density of training instances per class in the CamVid annotation, a common issue reported in the recent semantic segmentation literature. Our findings indicate a need for investigating class-balanced loss functions and instance-level data augmentation strategies.

DS-UNet-ResNet101 is presented in this work as a novel and powerful deep learning framework for high-precision semantic segmentation of urban road scenes. By combining a pretrained ResNet101 encoder with a symmetrical U-Net-style decoder and a multi-stage auxiliary supervision scheme, the proposed network mitigates two inherent limitations in conventional semantic segmentation: the conflict between semantic context and spatial details, and gradient degradation in deep network architectures. The synergistic combination of residual learning, skip-connected feature merging, and auxiliary supervision at different stages leads to a balanced and efficient learning process across both feature representation and spatial reconstruction. The use of ResNet101 as an encoder ensures the network can extract deep, hierarchical features that capture complex visual patterns across multiple scales in highly complex, variable urban scenes. These visual patterns are essential for representing the diverse, dynamically changing elements in urban road scenes, including roads, vehicles, pedestrians, and buildings and vegetation. The stacked decoder, composed of different progressively upsampling units in the spirit of U-Net, has enabled sufficient spatial resolution recovery during successive upsampling processes. At the same time, the skip connections from the encoder have played a decisive role in integrating low-level spatial details that were lost during encoding, leading to superior boundary-level classification performance, crucial in applications such as autonomous vehicles and smart traffic management. One of the most important contributions in this research is the introduction of a multi-stage auxiliary supervision technique.

The addition of auxiliary loss functions at different decoder layers enables gradient backpropagation to earlier network layers, alleviating the vanishing gradient problem and facilitating the learning of semantically rich intermediate representations. The weighting of the main loss and auxiliary losses further improved feature learning, leading to better segmentation performance across all classes. In practice, it was also observed that the loss curves with auxiliary supervision converge more smoothly than those of the single-loss network. The proposed DS-UNet-ResNet101 was evaluated on CamVid, a commonly used benchmark for urban scene understanding that comprises 32 object classes, various weather conditions, and a wide variety of scenes. Our model achieves an mIoU of 0.749, mPA of 0.755, and MAE of 0.068, outperforming a selection of well-established state-of-the-art networks, including FCN-8s, SegNet, U-Net and DeepLabV3+. The resulting 26-million-parameter network achieved excellent performance and quite good efficiency, considering both computational cost and segmentation accuracy. Qualitative examination of the segmentation outcomes also corroborates the effectiveness of our method, particularly for large objects with clear boundaries such as roads or vegetation. Similar to all methods that use an encoder-decoder architecture without an additional explicit attention mechanism, the DS-UNet-ResNet101 struggles to classify small objects with few pixels and those object classes characterised by high variability, such as pedestrians and cyclists, although the resulting segments remain plausible.

An additional contribution of this work is the deployment of the trained model into an interactive web application built with Flask. The application can perform inference in real time and present the resulting segmentation mask almost immediately after an image upload. The application furthermore allows browsing different datasets, visualising augmented samples and monitoring the training process. This practical example demonstrates the feasibility of applying the trained network and emphasises its potential for applications such as intelligent transportation and pedestrian and vehicle detection. In the context of this research, several interesting avenues for future exploration have been identified. It would be beneficial to extend the evaluation of our model to other public datasets with greater diversity in both scene and object classes, such as Cityscapes, KITTI, or ApolloScape, to assess the model's generalisation capabilities across different domains. The temporal information in the video streams could also be utilised via video segmentation methods (such as RNNs or 3D CNNs) to improve segmentation accuracy and stability. Also, an explicit attention mechanism in the channel or spatial dimension would lead to improved results, since the model can effectively capture long-range dependencies in the scene. Research might also focus on making the network efficient enough to run on resource-constrained edge devices. Model compression techniques such as pruning, knowledge distillation, or quantisation-aware training would enable our model to be deployed on resource-constrained devices. Additional solutions to address class imbalance and accurately localise small objects, such as custom loss functions or more effective data augmentation policies, would also be highly desirable.

To summarise, the proposed DS-UNet-ResNet101 network presents a well-balanced and highly effective framework for dense semantic segmentation of urban scenes, integrating the strengths of deep residual learning, an encoder-decoder architecture and a multi-stage auxiliary supervision strategy. The promising performance, high accuracy, and successful practical deployment suggest that it has great potential for various real-world applications in autonomous navigation and intelligent transportation systems, and serves as a robust foundation for future research on highly efficient, contextually aware segmentation models capable of handling an ever-increasing variety of dynamic and challenging scenes. The implications of the proposed DS-UNet-ResNet101 architecture extend far beyond the scope of this work and offer insights for a variety of interdisciplinary fields where precise, accurate scene parsing is vital. For example, in the context of intelligent transport systems, highly precise semantic segmentation can lead to higher levels of safety and situation awareness in both autonomous and semi-autonomous vehicles, and can also be integrated into general traffic management systems to enable more efficient regulation of traffic flows and to identify anomalous behaviours. These methods can be further employed to aid urban planning or to enrich surveillance systems for people counting and risk detection in public areas.

The versatility of the proposed architecture makes it applicable to other fields, such as medical imaging, remote sensing, and robotics, for dense pixel-level annotation and semantic analysis. In particular, this research highlights how to achieve a significant performance boost without the cost of a substantial increase in network size by carefully selecting the architecture's underlying components and the training strategy. The proposed architecture combines deep learning modules (deep residual learning) with traditional architectural design principles (encoder-decoder). These elements combine to leverage the properties of both methods to fully control the semantic and spatial information in complex scenes. From a methodological perspective, this study establishes a sound prototype that other researchers can use as a reference for future research on semantic segmentation, as the employed multidimensional testing approach provides a comprehensive view of the model's behaviour. This multidimensional evaluation can demonstrate validity not only for quantitative gains but also for visible and functional improvements. Future researchers can utilise this framework to ensure the proposed model is feasible across different situations. Besides the framework, the proposed model has been well integrated into the deployment platform, underscoring the importance of accessibility.

5. Conclusion and Future Work

As deep learning techniques are developing rapidly, the boundary between research and implementation must be narrowed down. In this study, a Flask-based interface not only demonstrated how advanced AI could be deployed on a convenient platform for end users but also enabled access to a powerful AI system with minimal computer knowledge, indicating that this trend may bring more AI toward democratisation. At the end, the research domain is continuously growing, and novel architectures, training methods and available datasets will be developed rapidly. The model based on the proposed architecture of DS-UNet-ResNet101 will continue to evolve and can incorporate the latest developments, including self-supervised learning, foundation models and multimodal data fusion, in future studies. By fusing more types of information (such as depth maps, LiDAR, and temporal video) into the learning process, the proposed model can perform even more accurately and robustly. Continuing exploration and research on hybrid architecture and adaptive learning strategy is expected to break the barriers in dense scene understanding. To conclude, this work contributes to the field as a whole, far beyond the immediate and specific gains on the CamVid dataset. This study presents a well-developed approach and technique that are feasible, efficient, and effective, and can benefit both research communities and industry.

Acknowledgement: The authors sincerely thank SRM Institute of Science and Technology at Ramapuram and Queen's University Belfast for their valuable support and encouragement in carrying out this research.

Data Availability Statement: The CamVid dataset used in this research is readily available on the official Cambridge-driving Labeled Video Database repository. The processed experimental recordings, trained DS-UNet-ResNet101 model weights and inference pipeline code which were developed in this research will be provided if reasonable requests are sent to the author in accordance with the current university data policy. All source code will be made available on GitHub upon final publication of the manuscript, for the model, pipeline, and Flask app.

Funding Statement: No specific funding grant was awarded for this work by any public, commercial or non-profit agency. This work was carried out as part of the undergraduate research. GPU hardware, high performance computer resources were provided by the institution.

Conflicts of Interest Statement: The authors report there are no competing interests of any nature. There are no affiliations of authors with other individuals or organizations with interests in the study. No financial, professional or personal relationships exist among the authors and any third party that can affect or unduly influence the results, presentation, or conduct of this study. All authors are fully aware of and agree with the version of the paper sent for publication.

Ethics and Consent Statement: The described work is based on computer analysis of publicly accessible road scene images extracted from the CamVid dataset. The CamVid dataset consists of video frames obtained from images captured in public traffic areas. As no human subjects were recruited, no personally identifiable information, nor sensitive personal data, were collected, nor was human or animal subjects subjected to any intervention or experimental manipulation. Accordingly, institutional board ethical approval was not required for this research. The dataset resources were used with full compliance of the terms and conditions specified in the public release agreement for the dataset.

References

1. S. Fan, J. Xie, Y. Zhuang, H. Chen, Z. Su, and L. Li, "BD-WNet: Boundary decoupling-based W-shape network for road segmentation in optical remote sensing imagery," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 18, no. 5, pp. 15523–15536, 2025.
2. C. Wang, X. Chen, B. Wang, L. Zhang, and B. Liu, "SLTM network: Efficient application of lightweight image segmentation technology in detecting drivable areas for unmanned line-marking machines," *IEEE Access*, vol. 12, no. 5, pp. 169001–169012, 2024.
3. Z. Ren, L. Wang, T. Song, Y. Li, J. Zhang, and F. Zhao, "Enhancing road scene segmentation with an optimized DeepLabV3+," *IEEE Access*, vol. 12, no. 12, pp. 197748–197765, 2024.
4. Y. Zhang, H. Song, Q. Wang, P. Jin, and T. Shen, "Semantic co-occurrence and relationship modeling for remote sensing image segmentation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 18, no. 2, pp. 6630–6640, 2025.
5. G. Luo, H. Li, J. Li, Z. Yin, and H. Wu, "SDCnet: A deep learning network model for road segmentation based on CNN and transformer," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 18, no. 9, pp. 23356–23371, 2025.
6. W. Ci, J. Xuan, R. Lin, and S. Lu, "A novel method for road anomaly objects detection in the traffic environment with multi-mechanism fusion," *IEEE Access*, vol. 12, no. 2, pp. 28369–28381, 2024.
7. A. Hoyos and M. Rivera, "Hadamard layer for improve semantic segmentation," *IEEE Access*, vol. 12, no. 12, pp. 194367–194377, 2024.
8. A. Rahi, O. Elgeoushy, S. H. Syed, H. El-Mounayri, H. Wasfy, and T. Wasfy, "Deep semantic segmentation for identifying traversable terrain in off-road autonomous driving," *IEEE Access*, vol. 12, no. 11, pp. 162977–162989, 2024.
9. Y. Yu, F. Peng, M. Zheng, and G. Yu, "SAM-guided state space model feature enhance network for remote sensing segmentation," *IEEE Access*, vol. 13, no. 11, pp. 198849–198863, 2025.
10. L. Nanni, S. Brahnam, and A. Loreggia, "An enhanced loss function for semantic road segmentation in remote sensing images," *IEEE Access*, vol. 12, no. 5, pp. 74218–74229, 2024.
11. Z. Li, Y. Yan, and B. Huang, "BuildWin-SAM: An improved SAM-based method for extracting building windows from street view images," *IEEE Access*, vol. 13, no. 4, pp. 61696–61707, 2025.
12. A. Manzoor, R. Mohandas, A. Scanlan, E. M. Grua, F. Collins, and G. Sistu, "A comparison of spherical neural networks for surround-view fisheye image semantic segmentation," *IEEE Open J. Veh. Technol.*, vol. 6, no. 2, pp. 717–740, 2025.
13. G. Liu, Y. Wang, Y. Zhou, S. Yao, and L. Han, "A residual complex Fourier for road segmentation in remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 18, no. 8, pp. 23936–23950, 2025.

14. J. Ren, L. Liu, Z. Xia, and Y. Liu, "AED-Net: A high-resolution remote sensing image road extraction method integrating atrous spatial pyramid pooling and efficient channel attention mechanism," *IEEE Access*, vol. 13, no. 3, pp. 61067–61077, 2025.
15. G. Kaur, S. Bharany, D. H. Elkamchouchi, and S. Kim, "Optimized ensemble learning for semantic segmentation of satellite imagery using DeepLabV3+ and UNet with PSO and cross-dataset evaluation," *IEEE Access*, vol. 13, no. 8, pp. 154647–154677, 2025.
16. X. Xiao, J. Tang, X. Lu, Z. Feng, and Y. Li, "A real-time road scene semantic segmentation model based on spatial context learning," *IEEE Access*, vol. 12, no. 11, pp. 178495–178506, 2024.
17. C. Chen, Y. Jia, L. Yu, and T. Liu, "A multi-scale context fusion network with boundary refinement for road marking segmentation," *IEEE Access*, vol. 14, no. 2, pp. 32221–32234, 2026.
18. H. Peng, S. Xiang, M. Chen, H. Li, and Q. Su, "DCN-Deeplabv3+: A novel road segmentation algorithm based on improved Deeplabv3+," *IEEE Access*, vol. 12, no. 6, pp. 87397–87406, 2024.
19. G. Ma, M. Zhang, J. Yang, Z. Shi, H. Ren, and Y. Zhang, "A label refining framework based on road matching and integration algorithm for road extraction," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, no. 10, pp. 19548–19564, 2024.
20. Y. Tang, P. Lyu, H. Shirai, Y. Watanabe, and H. Masui, "DroneRiver: An aerial semantic segmentation dataset for analysis of river leakage," *IEEE Access*, vol. 13, no. 8, pp. 137942–137951, 2025.

.

Publisher's Note: *The publisher remains impartial concerning jurisdictional claims in published maps and institutional affiliations. Responsibility for the content rests entirely with the authors and does not necessarily reflect the publisher's perspectives.*