Journal of Science and Transport Technology
Journal homepage: https://jstt.vn/index.php/en

Unravelling the Complexity of Indian Roads: Semantic Segmentation with LinkNet-UNet for Autonomous Vehicle Scene Understanding

Smita Khairnar[1,2], Suresh Kolekar[2], Shilpa Gite[2*], Biswajeet Pradhan[3*], Bhagyesha Patil[1], Shrutee Dahake[1], Radhika Gaikwad[1], Atharva Choudhri[1]
[1]Department of Computer Engineering, Pimpri Chinchwad College of Engineering, Nigdi, Pune 411004, India
[2]Symboisis Centre of Applied AI (SCAAI), Symboisis International (Deemed) University, Pune 412115, India
[3]Centre for Advanced Modelling and Geospatial Information Systems (CAMGIS), School of Civil and Environmental Engineering, Faculty of Engineering and IT, University of Technology Sydney, NSW 2007, Australia

Article info
Type of article:
Original research paper

DOI:
https://doi.org/10.58845/jstt.utt.2025.en.5.4.47-62

***Corresponding author:**
Email address:
Shilpa.gite@sitpune.edu.in
Biswajeet.Pradhan@uts.edu.au

Received: 28/06/2025
Received in Revised Form: 15/09/2025
Accepted: 07/11/2025

Abstract: As autonomous vehicles navigate through complex environments, understanding the scene poses a significant challenge. Traditional computer vision-based methods struggle to segment complex driving scenarios, making deep learning techniques increasingly popular. Semantic segmentation is essential for scene understanding in India's cluttered and diverse roads, where structure is often lacking. In this paper, we evaluate the performance of deep learning-based architectures including Linknet and Unet for semantic segmentation using the driving dataset of India's variety and cluttered roads, IDD-Lite. Through a series of experiments, we examine the combined effect of Linknet and Unet on IDD-lite. By harnessing the localization prowess and contextual understanding inherent in both models, our unified approach raises the bar for scene recognition tasks. Our approach is designed to work on Indian roads and prioritizes precision, efficiency and adaptability to various environmental conditions. Experimental results show our ensemble model has a MIoU of 0.69 and F1 score of 0.9. This is better than conventional ensemble methods and a big jump forward for semantic segmentation in autonomous driving systems for Indian roads.

Keywords: encoder-decoder, ensemble models, Indian Driving Dataset Lite, LinkNet, Semantic segmentation, Unet.

1. Introduction

Indian urban traffic is tough for autonomous driving systems. The roads are crowded, with multiple types of vehicles and inconsistent lane usage and traffic flow. To tackle these complexities, the Indian Driving Dataset (IDD) Lite was created to mimic the unstructured and densely populated road conditions of Indian cities [1]. As self-driving technology evolves to handle such environments, high precision scene interpretation becomes more important. Traditional road segmentation techniques fail on these roads which

are irregular and chaotic. Deep learning methods which rely on large and diverse datasets for training are more adaptable. Datasets like IDD Lite enable these models to learn and respond to the variability in Indian traffic scenarios [2].

Semantic segmentation is key for autonomous cars to understand driving scenarios. Precise mapping of category labels to image pixels allows to identify important features like roads, lanes, cars, signs and people. This pixel level parsing helps in navigation planning and decision making. Scene complexity, occlusion, size variation, and other factors make segmentation difficult [3]. Crowded roads, lack of discipline among road users, diverse traffic patterns, varying illumination and weather conditions. Addressing these specific challenges necessitates tailored solutions. The Indian Driving Dataset (IDD) Lite benchmark covers dense metropolitan areas and highways, including various scenarios. Until recently, segmentation algorithms depended on hand-crafted features; however, with enough training data, deep neural networks currently dominate. Encoder-decoder designs such as LinkNet provide efficiency by progressive downsampling and upsampling, balanced with lateral connections in UNet, resulting in a multiscale context. Such hybrid models with pretrained encoders attain great accuracy, as LinkNet and UNet models demonstrate.

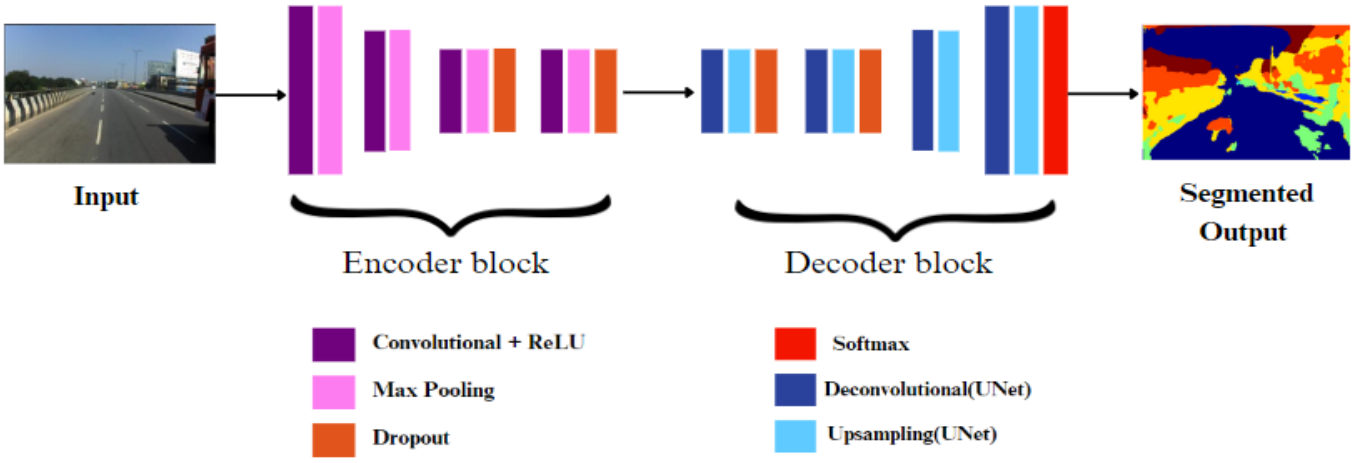


Fig. 1. Layout of an encoder and decoder based on Semantic segmentation models

UNet developed an encoder-decoder architecture with connections that bypass segmentation, allowing for precise localization and various scale contexts. The encoder gradually downsamples input pictures to record high-level context, whilst the symmetrical decoder increases to full fidelity to produce pixel-accurate results. LinkNet improves the fundamental encoder-decoder structure's efficiency by requiring only a few selected skip connections rather than complete concatenation [4].

The LinkNet and UNet models are highly suitable for the image segmentation of the IDD-Lite (Indian driving dataset) as they balance the contextual representation and accuracy. While the encoder-decoder structure of LinkNet allows for fine spatial localization, the skip connections of UNet exhibit multiscale properties. Together, they provide a lightweight and accurate framework for parsing complicated driving scenarios. Studies conducted on the Indian Driving Dataset Lite support these mutually beneficial impacts, with the combined model beating any individual model. The combined framework learns category-level characteristics and excellent spatial cues required for detecting critical items, such as road agents, during navigation in congested urban environments. Overall, the model's ability to quickly and accurately encode spatial features, adaptability via the encoder-decoder framework and use of local and global context makes it a good fit for Indian road conditions in the IDD Lite dataset.

We are driven by the need for specialized metrics and algorithms to extract structure from the chaos of Indian roads. Fig. 1 shows the encoder and decoder architecture for semantic segmentation tasks. These designs use many layers that build receptive fields to record low- and high-level information.

This paper makes several key contributions to the field. Firstly, it recommends the adoption of the LinkNet-UNet paradigm for semantic segmentation within India's unique environment, specifically focusing on its application in autonomous vehicle operations. Secondly, it meticulously evaluates and scrutinizes the effectiveness of the fused LinkNet-UNet model compared to other state-of-the-art deep neural network semantic segmentation models. Lastly, it validates the performance of this proposed fusion approach by leveraging the benchmark IDD-Lite dataset, demonstrating its efficacy in handling the complexities of real-world Indian driving scenarios. Overall, these contributions collectively enhance our understanding and application of semantic segmentation techniques within the context of autonomous vehicle navigation in challenging environments like India.

Our new approach is based on the fusion of UNet with the LinkNet model. This generates a synthesis of the two models' abilities. This synergistic combination can use UNet’s Semantic segmentation and Linknet's very effective sampling. In comparison to using each model on its own, the results of these experiments demonstrate that our fusion technique is efficient regarding linguistic segmentation challenges and shows significant improvements in accuracy and MIoU. We found that the accuracy was 0.9039 when we combined these two models and compared them. The validation MIou for the same is 0.6926. The results of this study show that combining the UNet and LinkNet models will produce an optimal blend that exceeds the different models. This demonstrates the value of integrating different skill sets into a comprehensive approach.

The paper is organized as follows: Section 2 covers relevant research on autonomous cars and semantic segmentation. Methodology for the proposed architectural system is detailed in Section 3. Section 4 discusses the Indian Driving Dataset (IDD-Lite) and its experimental setup. An overview of experiments and discussions is presented in Section 5. Section 6 summarizes completed research and results. References are listed in Section 7.

2. Literature Review

Semantic segmentation is a fundamental problem in computer vision. It involves grouping an image into semantically meaningful regions and labelling each pixel to the object or class it belongs to. Over the past decade, significant progress has been made in semantic segmentation mainly due to the advent of deep learning techniques, especially convolutional neural network models (CNNs). This literature survey aims to provide a comprehensive overview of recent developments and techniques for semantic segmentation. The papers below explore key concepts, challenges, and trends in the area to understand the current state and future directions for semantic segmentation research.

In [1], they explain their research to increase public trust in self driving cars by listing the performance metrics to evaluate the GradCAM Inception UNet model to see a scene. Evaluation of the model is to look at metrics like recall, accuracy and precision or similar. To show how the Inception UNet model understands driving conditions, this paper uses a gradient weighted class activation mapping algorithm called GradCAM. Within structural segmentation, GradCAM looks at which parts like pedestrians affect the model’s decision making and categorizes them accordingly.

Deep learning methods like DFE-AVD enhance vehicle detection by combining multiple CNN models through ensemble techniques [5]. Autonomous vehicles rely on deep learning for environmental perception, handling challenges

such as lighting, weather, and road conditions using CNNs trained on large datasets [6]. Similarly, UAV-based remote sensing leverages CNNs for land classification and object detection, addressing the scarcity of high-quality labeled data [7].

The autonomous driving system needs to model and forecast dynamic agents like vehicles and pedestrians. Artificial intelligence techniques like machine learning and computer vision are used for that. When understanding complex traffic patterns, data-driven approaches like neural and Bayesian networks work. Further research is necessary in uncertainty management and mixed artificial intelligence [8]. Deep neural networks are being used to train control policies for self-driving cars, with recent surveys analyzing DRL approaches for vehicle control, path planning, and navigation tasks. While promising for simulation and closed-track applications, deep learning remains a challenge to ensure safety and clarity in the decision-making process when driving on public roads [9].

In [10], the authors studied integrated architecture frameworks for context modelling using neural modules and graphical models, focusing on human-centric mechanisms, causality, spatial-temporal dependencies, and rules, but faced challenges in commonsense reasoning, model decision explanation, and safety compliance. They dealt with the issue of the prediction of motion in autonomous cars, which is crucial for navigation and route planning. They used a massive amount of a realistic Lyft Level 5 prediction dataset containing over 55,000 real-world driving events. Further, they presented a deep neural network model based on the CNN LSTM encoder-decoder architecture to predict the vehicle's trajectory.

In [11], the model incorporates dynamic and static scene context, such as traffic signals or prior movements. Their study revealed that a significant contributor to road safety concerns is the increased usage of motor vehicles without improvements to safer systems, highlighting the need for priority action in areas like speed management and safety enforcement. In [12], the authors proposed LaneNet, a specialized encoder-decoder CNN for automated drive perception, aiming for 10ms inference speed on a single GPU. It prioritizes the early output of lane-related semantic features, optimizing runtime performance without sacrificing quality. Trained on TuSimple datasets, LaneNet demonstrates state-of-the-art FPS performance with strong accuracy metrics. Open-sourcing LaneNet and evaluation tools push the boundaries of efficiency-aware algorithm design [13].

Deoli et al. (2024) studied off-road automated vehicle segmentation models' vulnerability against adversarially manipulated inputs [14]. They created a benchmark dataset that shows weak robustness and proposed an analytical training methodology in multiple scales of antagonistic training for external robustness. Xu et al. (2023) compared the optimum selection of loss functions for deep neural network-based road segmentation models using overhead remote sensing imagery. They compared eight loss formulations with two publicly available data sets and found substantial deviations in performance. Optimized hybrid Dice+cross entropy losses are identified as the ideal solution to the problem of class imbalance. These findings are of interest to the remote sensing community [15].

[16] developed a unique convolutional neural network architecture called Eff-Unet, which is specifically designed to process scenes captured during rapid driving. This architecture has used the UNet structure to combine contextual data with complex geographic details by combining encoder-decoder skip links. A significant advance has been made in adding effective attention management blocks. These blocks help the network focus on the primary components by weighting map features in relation to them. The authors highlighted the trade-offs made in accuracy, inference time, and model size by contrasting Eff-Unet with other segmentation CNNs currently in use. To assess the effectiveness of EffUNET in realistic scenarios,

data from India and Cameroon have been used to evaluate it based on datasets that show unpredictable driving conditions. The project's objective was to use a robust semantic segmentation approach to solve specific problems associated with unconcealed driving situations [16].

To address the challenges of generalizing across various road textures, lighting conditions, vehicle types, and cluttered backgrounds encountered in uncontrolled driving environments, [4] developed an ensemble model based on meta-learning. This approach aimed to enhance the model's capability to handle unexpected variations in real-world driving situations. The proposed NuWaNet architecture utilized an omniscient supervised training method, which combined labeled samples, synthetic labels, and independently supervised auxiliary tasks to optimize the diversity of the training data. Combining predictions from multiple NuWaNet variations through trained fusion techniques, the final segmentation result was obtained during meta testing. Tests conducted on unstructured data from India and Malaysia demonstrated greater flexibility than previous approaches. The study provides a valuable foundation for comprehending complex, highly heterogeneous scenarios in real-world settings.

[17] introduces a novel contextual data aggregation module named OCNET. This module systematically aggregates multilevel features under an umbrella pathway to enhance segmentation in complex scenarios. The authors have devised a method to accurately and efficiently interpret scenes, utilizing a dual-path coordination system incorporating compact decoders and ResNet encoders with Layerwise Interoperability. SegFormer, a novel method for semantic segmentation that utilizes transformers instead of convolutional neural networks (CNNs), has been developed [18]. Their model uses a meticulously crafted transformer encoder to attain comparable performance to CNN techniques, eliminating the requirement for convolutional layers. This innovation represents a significant shift in the semantic segmentation domain.

In [19], a Pixelwise encoding Network was introduced for scene segmentation. This paper defines the challenge of scene segmentation, a critical task for autonomous vehicles, as a classification difficulty [19]. In another study, they incorporated attenuating convolutions into the FCN-VGG16 model to improve road scene segmentation. Comparing this approach to previous ones, the mean Intersection over Union (IoU) on the CamVid test set significantly increased to 73.1% [20].

To perform natural scene segmentation, [21] integrated VGG19 with fully linked CRFs, achieving a 79.7% MIoU on the Pascal VOC 2012 benchmark. [22] discusses TernausNet, which utilizes pre-trained VGG19 encoders and decoders with U-Nets in between, achieving 84.5% accuracy on the ISBI Challenge for segmentation. DeepLab v3+, based on the Xception model, and incorporated depth-wise separable convolution built on top of ResNet. They achieved an MIoU of 82.1% on the Pascal VOC 2012 leaderboard [23]. In another study, [24] optimized a RefineNet for cardiac MRI segmentation using ResNet-152 as an encoder, achieving a dice metric of 0.938 on the ACDC dataset.

In [25], the authors utilized a UNet segmented framework with a MobileNet v2 encoder pre-trained on ImageNet for satellite imagery, demonstrating a good balance between accuracy and efficiency. They also developed a real-time scene parsing network using a modified MobileNetv2 encoder and efficient propagation of space modules, achieving 71.4% mIoU on cityscapes at 67 FPS inference speed. The study enhanced LinkNet by incorporating multiple scales feature aggregating interconnections for robust segmentation of irregular Indian driving datasets, achieving a mean IoU of 66.2% on the IDD challenge. In contrast, the study in [26] explores the UNet model with ResNeXt encoders to address

low-light, transient camera captures in feature-sparse traffic monitoring datasets, achieving 87.3% accuracy for nighttime road scene segmentation.

Various deep learning techniques, including architectures like VGG16 and VGG19, have been employed for different applications, such as face liveness detection, as demonstrated in [27]. They have also been utilized for Iris liveness detection [28]. Additionally, researchers [29] have utilized transfer learning methods like AlexNet, VGG16, and VGG19 with pretrained versions for analyzing species of Pistachio, resulting in higher success rates. Evaluation was conducted on specific parameters such as precision, F1 Score, specificity, and sensitivity. The most recent research [30] involves using various deep-learning models for image style transfer to extract features and generate images with artistic value.

While existing approaches have produced outstanding results, there is always room for future performance improvements. This paper proposes merging chosen models as a possibility of advancement. Fusing techniques and measuring the mean Intersection over Union (mIoU) for these models on benchmark datasets will determine if our proposed variants can improve state-of-the-art accuracy. Additional tests combining effective strategies may uncover complementary strengths and accelerate advancement beyond existing segmentation capabilities.

3. Methodology

Various network topologies create exact segmentation maps on complicated real-world pictures motivated by data sets, metrics, and scene interpretation issues. The proposed technique involves experimenting with several models, such as LinkNet and Unet. Fusion of two models is also performed, specifically LinkNet and UNet.

LinkNet and UNet are encoder-decoder convolutional neural networks used for semantic segmentation. Both designs include an encoding system that extracts hierarchical characteristics from an input picture using techniques such as convolution and pooling. The primary distinction is the lack of connections between the encoder and decoder. In LinkNet, the encoder output is linked directly to the decoder via skip connections, which transport information from previous encoder layers. The decoding device recovers spatial information and uses the conveyed data to create precise pixel-wise segmentation masks. UNet uses skip connections, but only between matched encoder and decoder layers. Lower-level encoder characteristics are transferred straight to many higher-resolution decoding layers during upsampling. This multiscale synthesis of information and context allows for exact localization in the final segmentation masks.

Combining the respective encoder-decoder modules benefits the fusion approach from the complementary characteristics of the LinkNet and the UNet designs. However, LinkNet can quickly develop accurate Semantic Maps, even if it often suffers from the distortion of object borders. As a result, UNet can precisely identify complicated shapes at significantly higher computational costs than other methods. Fusing them enhances their complementary exact positioning from Unet and performance from LinkNet while mitigating their limits. By combining these two, fusion model improves the segmentation results by speeding up and improving the accuracy of LinkNet and UNet.

The feature maps from the last convolutional layers of the encoders of both Linknet and Unet models of the same size are fused by making sure the dimension of the individual feature map and the fusion feature map are the same. The output of these fused layers are then passed as input to the decoders of both models separately. Further the last convolutional layers of LinkNet and Unet are merged, we get the final output as the segmented image. So combining LinkNet and Unet in image segmentation is a strong solution for applications that requires both speed and accuracy in segmentation tasks.

Fig. 2 depicts the suggested system combining Unet and LinkNet architectures to perform image segmentation tasks. To extract a

feature map, an input image with dimensions of 256x320x3 height, width, and color channels is fed into the system and processed by the Unet Encoder 256x320x7. In addition, two different decoders called LinkNet256x320x7 and Unet259x320x7 are input to the feature map. Both decoders are upgrading the feature map to the original input resolution. Combining the final convolution layers of UNET and LinkNet decoders gives you a single output. The output of the LinkNet decoder is converted to a fused output from both designs' final convolution layers in an ensuing transformation step. In this final fusion phase, the features of the Unet and LinkNet architectures are combined to create a segmented output image of 256x320x7. In the segmented output image, the input image is divided into pixels, and each pixel is assigned a class label based on the attributes taught to it by the UNet and LinkNet designs. By exploiting the advantages of both architectures, this fusion method aims to improve the segmentation of images.

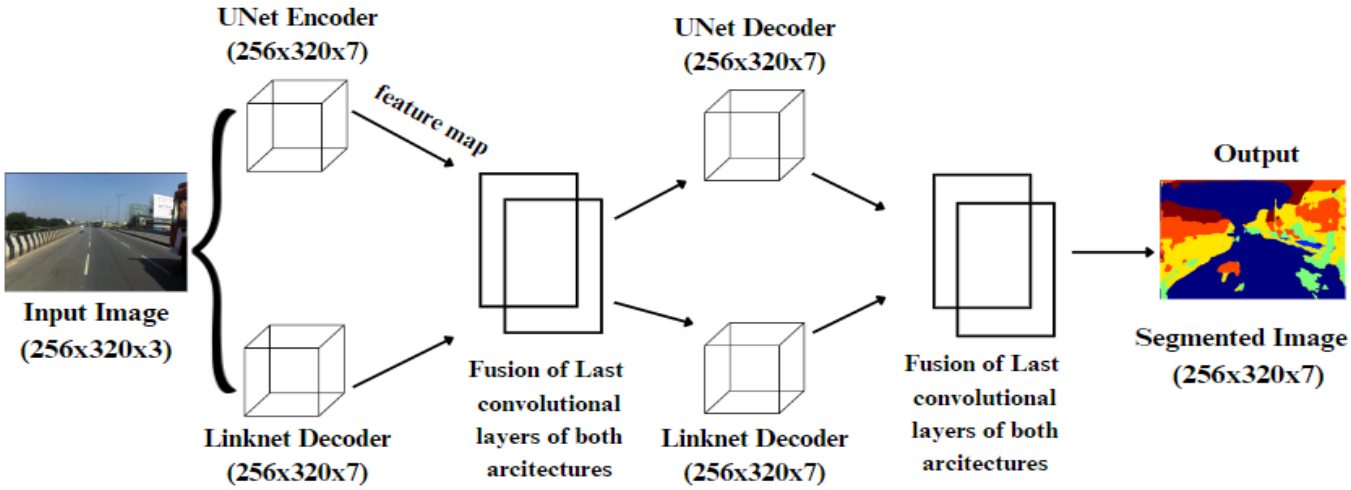


Fig. 2. Block diagram of the proposed system using a fusion of Unet and LinkNet for image segmentation

4. Experimental Setup

This section includes the dataset, machine, configuration, and performance measures.

4.1. Dataset

Fig. 3 demonstrates IDD Lite (Indian Driving Dataset), a semantic segmentation dataset focusing on Indian driving situations. These images depict a range of driving settings, including low, medium, and congested traffic, highways, country roads, and even nighttime driving. Roads, vehicles (cars, trucks, etc.), two-wheelers (bikes, scooters), roadsides (people, excluding vehicles), sky, and buildings are among the key categories marked. There are also some miscellaneous categories. This dataset is around 50 MB and has 1380 training pictures and 204 test images at a resolution of 320 × 227 pixels. It comprises two files: leftImg8bit and gtFine, which contain labels for leftImg8bit. gtFine has the val and train folders, whereas leftImg8bit contains the val, train, and test folders.

IDD Lite's diversity of sceneries, with multiple dynamic actors and varied road kinds, makes it difficult for segmentation algorithms to test the limits of current methodologies. It has enabled benchmarking semantically segmented CNN models designed for self-driving visual processing in challenging urban situations [12].

Fig. 4 offers visuals and annotations for seven distinct items typically encountered in driving situations. The courses are living things, vehicles, non-drivable areas, sky, roadside, and far objects. These classes have been marked in the IDD Lite dataset to help train and evaluate object identification techniques for driving scenarios.

Fig. 3. Sample images from IDD-Lite Dataset

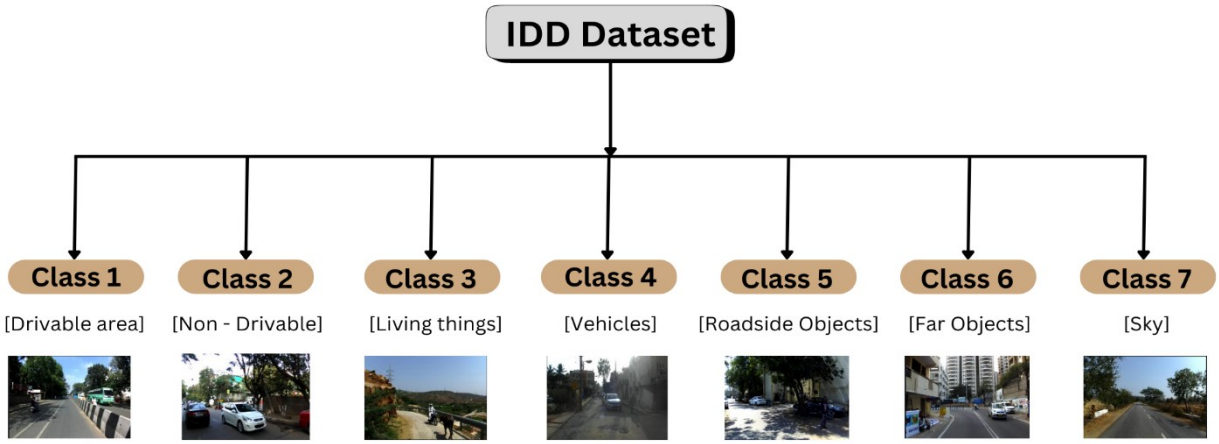


Fig. 4. Representation of IDD-Lite Dataset with 7 classes

4.2. Model Training

The model was developed and trained using a system equipped with a Windows-based operating environment, powered by an AMD Ryzen 5 processor and 8 GB of RAM. Google Colab Pro was utilized for enhanced computational capacity, offering a 64-bit architecture, 51 GB of RAM, and GPU support. The training and testing of the model were executed within the Colab Pro platform, which provided access to an NVIDIA T4 GPU and substantial memory resources suitable for deep learning tasks.

4.3. Performance Metrics

The model trained on the data for semantic

segmentation performed admirably, as evidenced by numerous essential measures. LinkNet and Unet employ performance measurements for individual models, including loss, accuracy, mean IoU, val_loss, val_meanIoU, and val_accuracy. For the fusion of models LinkNet and Unet, performance measures such as loss, accuracy, mean IoU, val_loss, val_meanIoU, val_accuracy, specificity, sensitivity, and F1 score are employed. Specificity, sensitivity, and the F1 score are standard assessment measures for semantic segmentation tasks.

Specificity refers to the fraction of accurate pessimistic predictions among all real negative samples.

It's computed as:

$$\text{Specificity} = \frac{True\ Negative}{True\ Negative + False\ Positive} \quad (1)$$

Where TN: (true negative) forecasts, whereas FP: (number of false optimistic predictions).

Sensitivity is the fraction of accurate optimistic predictions of all real positive samples. It's computed as:

$$\text{Sensitivity} = \frac{True\ \ Negative}{True\ Negative + False\ Positive} \quad (2)$$

Where TP denotes genuine optimistic predictions, and FN denotes false pessimistic predictions.

The F1 score is the harmonic mean of recall and precision, and it is a measure of a model's overall accuracy. It's computed as:

$$\text{F1 score} = 2 * \left(\frac{Precision * Recall}{Precision + Recall} \right) \quad (3)$$

Precision is defined as the fraction of genuine optimistic predictions (correctly predicted positive cases) among all instances projected as positive by the algorithm.

$$\text{Precision} = \frac{True\ \ Positive}{True\ Positive + False\ Positive} \quad (4)$$

MIoU (mean intersection over union) averages IoU values (Jaccard index) across all the masks. IoU is computed as:

$$\text{IoU} = \frac{True\ Positive}{True\ Positive + False\ Positive + False\ Negative} \quad (5)$$

The semantic differentiation measures for IDD Lite can be derived from comparing the predictions made by the model to the ground truth labels. Higher specificity, sensitivity, and F1 score indicate improved performance of the segmentation model. This showcases consistent segmentation accuracy across all classes, including pedestrians, cars, and roads. The model demonstrates efficient training for semantic segmentation on IDD Lite in a generalized setting, evidenced by excellent accuracy, region-based IoU metrics, and minimal loss values on training and validation datasets. Despite these achievements, there remains room for further advancement through alternative approaches.

5. Results

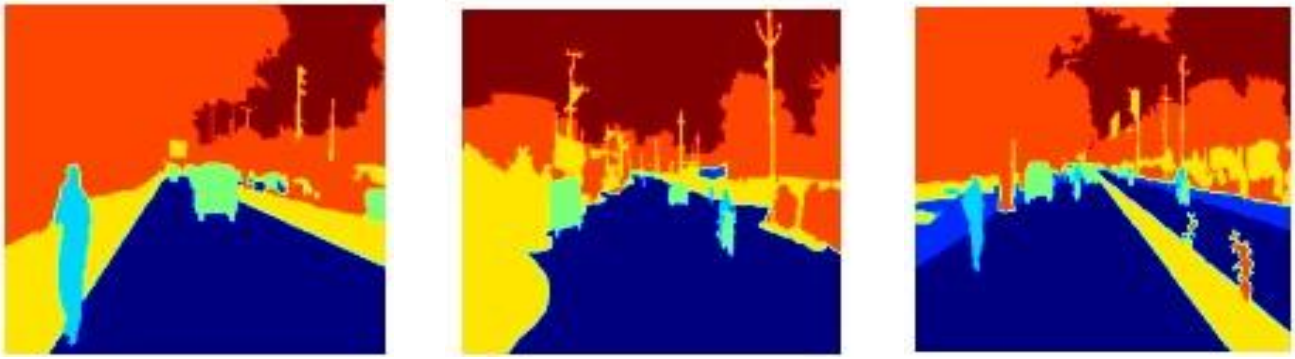
Using different models, such as LinkNet and Unet, the overall accuracy for 50 epochs was determined. It helps the model learn more accurate representations of features and patterns from the training dataset. The chaotic traffic patterns found in many developing world cities present distinct challenges for the perception systems of autonomous vehicles intended to navigate these unstructured environments safely. Unlike the organized flow of cars in structured urban settings, roads in these locations often have undefined lanes and streams of cars, bikes, pedestrians, carts, and animals intermingling unpredictably. Multi-label semantic segmentation algorithms that accurately categorize all road users in the scene and object detection networks pinpointing individual entities and their trajectories are critical front-end capabilities. Autonomous vehicles can only anticipate the actions of other drivers and create safe, defensive routes by detecting and understanding the chaos around them. However, progress in developing and evaluating perception pipelines specialized for disorderly developing

world driving has yet to be improved because of the lack of diverse public datasets. The Indian Driving Dataset (IDD-Lite), capturing hours of driving in complex urban and highway environments across Indian cities, remains the sole resource available to researchers to date.

Figs. 5, 6, and 7 show the input images, ground truth images, and their respective segmented output image of the proposed fusion architecture.

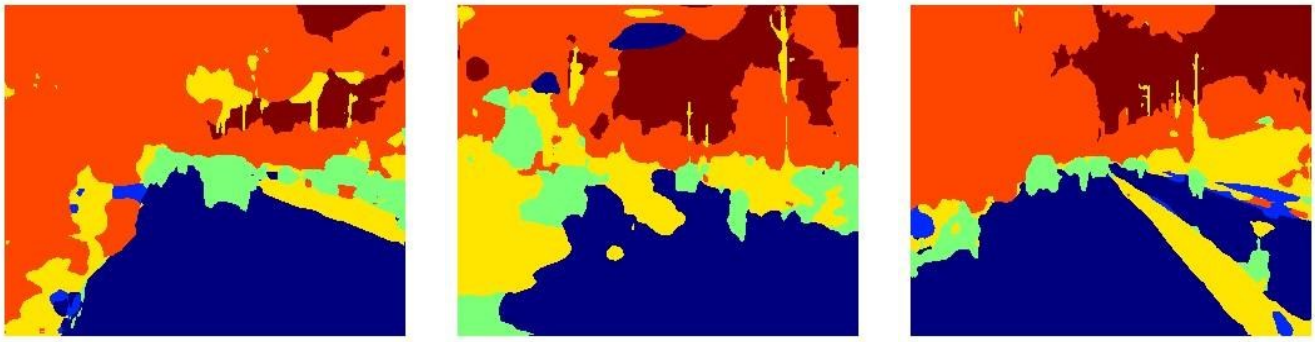
(a) (b) (c)

Fig. 5. Sample input images from the original IDD-Lite Dataset



(a) (b) (c)

Fig. 6. Ground truth images of respective sample input



(a) (b) (c)

Fig. 7. Segmented output of respective input images for the proposed model

Fig. 8 depicts values of performance metrics across 50 training epochs for each model tested. Accuracy measures total pixel-level accuracy in semantic segmentation predictions. The graphic allows you to evaluate convergence and stability as optimization progresses. Fig. 8a depicts each architecture's Mean Intersection over Union (MIoU) throughout 50 epochs. In addition to accuracy, MIoU is an important performance metric for semantic segmentation since it quantifies the class-specific gap between predictions and the truth on a per-class basis. Evaluating model sensitivity over time offers information on proper semantic class discovery consistency. Fig. 8b

shows the models' per-epoch accuracy values throughout 50 epochs. Higher numbers imply a better capacity to detect the existence of a class accurately.

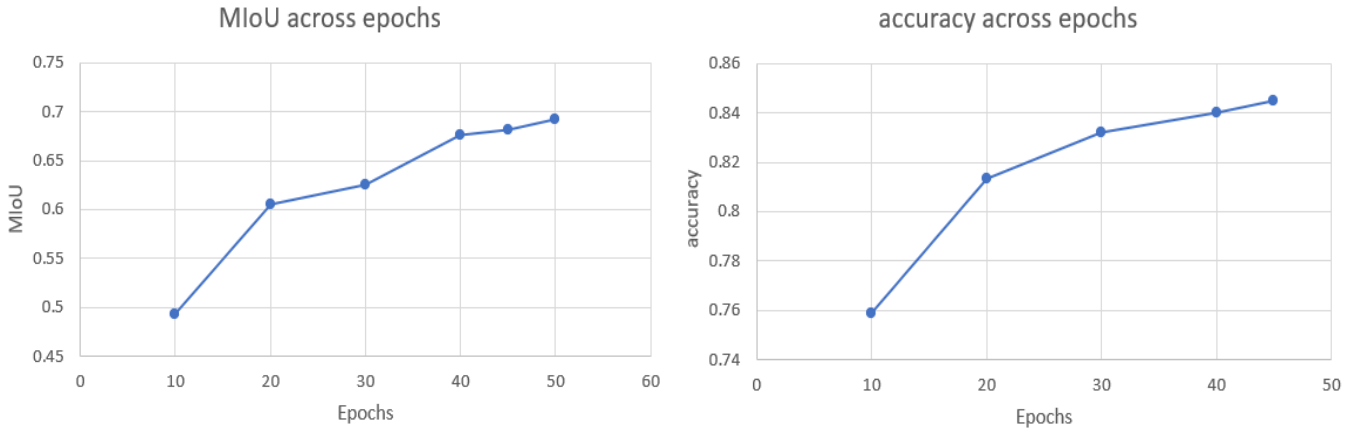
Finally, Fig. 8c shows the per-class F1 scores for each model during optimization. The F1 measure takes into account both accuracy and recall. Following F1 over-training examines balanced progress in accurately forecasting pixels of a particular class while collecting every pixel of that class. At the same time, Fig. 8d represents sensitivity across epochs. These four graphs, when combined, provide for a complete assessment of many elements of segmentation performance throughout each model's iterative optimization process.

Table 1 depicts the performance results for numerous current semantic segmentation models. Performance metrics such as Mean IoU, Accuracy, Specificity, Sensitivity, F1 Score, and Dice Coefficient are mentioned.

The results of the performance evaluation for a proposed fusion design combining LinkNet and UNET models are presented in Table 1. The fusion model is perfect in terms of training data and a number of metrics. There is a significant alignment between the model's predictions and the actual segmentation, as indicated by the mean Intersection over Union (IoU), a measure of the overlap between the ground truth and anticipated masks, which is reported at 0.7636. The model can accurately classify pixel with an accuracy score of 0.9039. In addition, the model's robustness and sensitivity values demonstrate that it can distinguish between actual positive and true negative situations. The F1 score, the harmonic mean of accuracy and recall at 0.9854, also highlights the balanced performance of fusion models in all classification elements. In addition, the Dice coefficient, measuring spatial overlap, is reported to be 0.8640, indicating a high level of agreement between segmentations that have been predicted and those based on factual evidence. As a matter of fact, these findings show that the fusion architecture can segment images with a high degree of precision, recall, and spatial agreement.

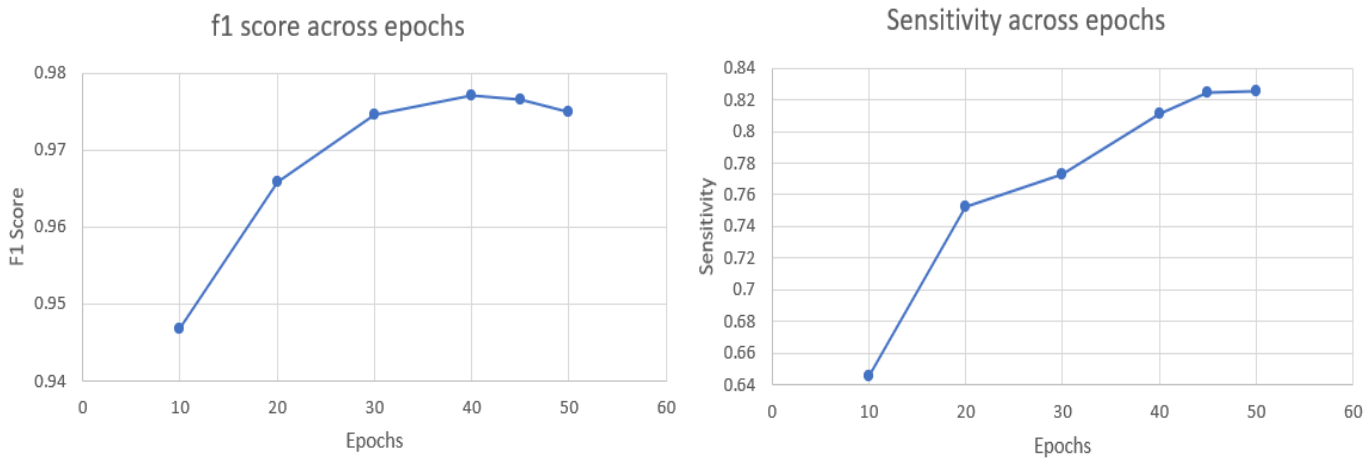
Table 2 displays the results of benchmarking experiments, which evaluated the effectiveness of existing semantic segmentation approaches to the proposed novel model. The results show that the suggested model outperforms existing cutting-edge techniques on important parameters such as mean IoU. This demonstrates the better segmentation capability provided by the new model.

The performance of existing popular semantic segmentation methods is compared in Table 3. This allows for a quantitative evaluation of the benefits and disadvantages of each recognised approach. The comparative study overviews the current environment and semantic segmentation capabilities.



(a) MioU values across epochs (b) Accuracy values across epochs

Fig. 8. performance metrics of proposed model across epochs



(c) f1 score values across epochs (d) Sensitivity values across epochs

Fig. 8. (continued)

Table 1. Performance evaluation for the proposed fusion architecture

Model	Dataset	Mean	Accuracy	Specificity	Sensitivity	F1	Dice
Fusion of Linknet and	Training	0.7636	0.9039	0.9888	0.8869	0.9854	0.8640

Table 2. Performance evaluation for the proposed fusion architecture

Model	Mean	Accuracy	Sensitivity	Specificit	F1 Score
E-Net	0.566	0.9321	0.8669	0.9395	0.7229
SegNet	0.3076	0.8971	0.8896	0.8975	0.4705
U-Net	0.6031	0.9203	0.8534	0.9500	0.7056
Inception U-Net	0.622	0.958	0.728	0.975	0.740
Proposed fusion model	0.6926	0.8457	0.8255	0.9804	0.9750

Table 3. Comparison of the proposed model with the SOTA models based on MIOU

Model	Proposed	U-Net	Inception U-	UNet-	SegNet	E-Net	UNet-
MIoU	0.6926	0.6031	0.622	0.5981	0.3076	0.566	0.6174

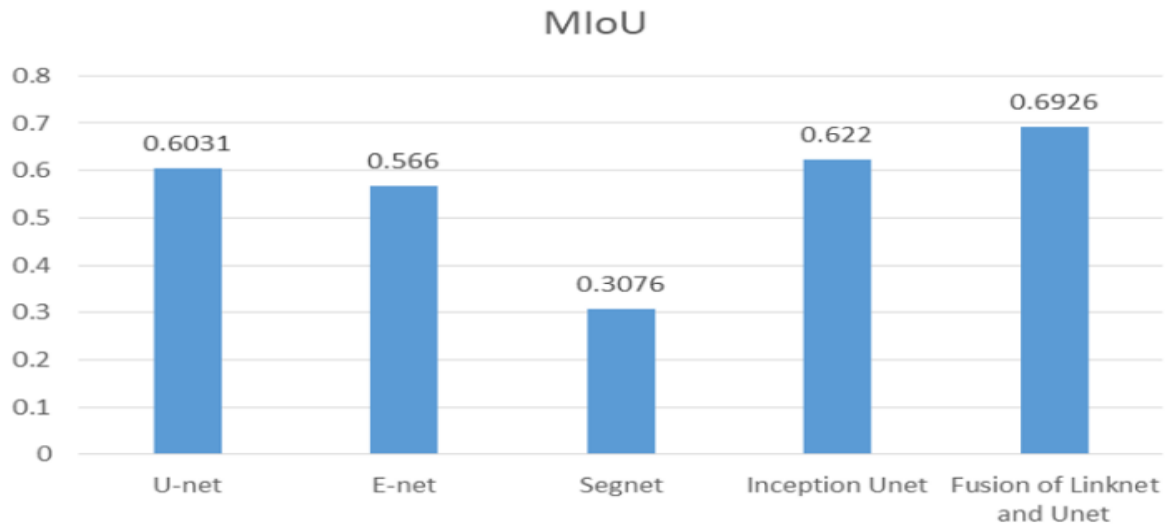


Fig. 9. Comparative analysis based on MIoU v of the proposed model with SOTA models

6. Discussion

Road scene segmentation is the backbone of Intelligent Transportation Systems (ITS) and is critical for navigation safety and situational awareness of autonomous vehicles. In densely populated and high traffic areas, the ability of these systems to understand their environment and predict the movement of nearby vehicles is key to performance. To facilitate research and innovation in this space, many industry leaders—Waymo, LYFT, Argoverse etc.—have made public datasets. But these datasets are mostly collected from highly structured traffic scenarios in developed countries. Addressing a different set of challenges, the Indian Driving Dataset (IDD) was created to represent the unstructured and less regulated road conditions of countries like India where traffic patterns are unstructured and unpredictable.

In this work, a custom hybrid segmentation model was built by combining U-Net and LinkNet architectures. The methodology involved combining the encoder and decoder outputs of both the networks into one framework. This dual fusion approach was designed to improve the segmentation accuracy by leveraging the strengths of both the models. Once the fusion model was built, it was compared with several well known semantic segmentation models.

The proposed fusion model performed well with an accuracy of 0.9039 and mIoU of 0.7636 on the training dataset. The proposed model performed better than other models on multiple metrics when validated against E-Net, SegNet, U-Net and Inception U-Net. Specifically, the validation mIoU was 0.6926 which is better than the other models which had mIoU of 0.566, 0.3076, 0.6031 and 0.622 respectively.

Beyond mIoU, the proposed model also led in specificity, precision, and F1 score. Precision scores for the other models ranged from 0.8975 to 0.975, whereas the proposed model achieved 0.9804. Similarly, in terms of F1 score, the baseline models scored between 0.4705 and 0.740, while the fusion model reached an impressive 0.9750. This notable performance boost is attributed to the novel fusion technique, combining the final convolutional layers of U-Net and LinkNet's encoder and decoder paths, enhancing feature representation and learning depth.

Performance metrics are detailed in Tables 1–3, while Fig. 9 visually compares model effectiveness based on the mIoU values.

One limitation of the fusion model of U-Net and Link-Net is its potential for increased computational complexity. While combining the strengths of both architectures can improve performance, the fusion may result in a model with a larger number of layers and parameters. This can increase the computational burden during training and inference, requiring more computational resources and time than using either U-Net or Link-Net individually. Therefore, while the fusion approach offers the promise of enhanced performance, it comes with the trade-off of heightened computational demands.

In the future, this can be improved by applying model optimization techniques to reduce the computation without sacrificing accuracy. Also, attention mechanisms or transformer modules can be added to improve segmentation quality especially in complex urban scenes. Domain adaptation techniques can also help to generalize the model across different driving environments.

7. Conclusions

Extensive research and assessment have shown that combining these models produces better outcomes than separate designs alone. The integrated model incorporates high-level context and fine-grained features, resulting in more robust and accurate scene interpretation in harsh driving conditions. Fusing the LinkNet encoder-decoder with UNet connections shows potential for precise semantic segmentation on the IDD Lite dataset. The fused model outperforms each architecture because the LinkNet model captures global context, and the UNet model correctly localizes classes. Our results show that overall accuracy outperforms the distinct LinkNet and UNet models

by 0.9039 each. Mean IoU improves by up to 0.6926, showing that the fused model is more reliable at classifying separate items such as vehicles and traffic signs. The fused decoder aids the model's generalization to new test data, as evidenced by higher validation metrics. These improvements can have practical applications in real-time driver assistance systems, especially in regions with unstructured traffic conditions. The model can also be used to enhance road infrastructure monitoring, helping urban planners identify traffic bottlenecks and areas needing safety interventions. There is room to improve these outcomes by including more data and attempting different fusion combinations. However, the work clearly shows the promise of blended decoder modules for better segmentation, with direct application to visual scene interpretation for autonomous car navigation.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Authors Contribution

Conceptualization: S.C.K. and S.D.T.; data preparation: B.P. and S.D.; writing original draft: R.G., A.C., B.P., and S.D.; supervision: S.C.K. and S.D.T.; methodology: S.D., B.P., R.G.; visualization: S.D. and B.P.; validation: S.C.K.; review and editing: S.C.K., R.G., A.C., B.P., and S.D.; project administration: S.C.K. and S.K.; resources: S.C.K., S.D.T., and B.P.

All authors have read and agreed to the published version of the manuscript.

Data Availability Statement

The data used in this study are available at https://idd.insaan.iiit.ac.in/dataset/details/, accessed in January 2024.

References

[1] S. Kolekar, S. Gite, B. Pradhan, A. Alamri. (2022). Explainable AI in scene understanding for autonomous vehicles in unstructured traffic environments on Indian roads using the inception U-Net Model with Grad-CAM visualization. *Sensors*, 22(24), 9677. https://doi.org/10.3390/s22249677

[2] T. Tiwari, M. Saraswat. (2023). A new modified UNET deep learning model for semantic segmentation. *Multimedia Tools and Applications*, 82, 3605-3625. https://doi.org/10.1007/s11042-022-13230-2

[3] D. Singh, A. Rahane, A. Mondal, A. Subramanian, C.V. Jawahar. (2022). Evaluation of Detection and Segmentation Tasks on Driving Datasets. *In: Raman, B., Murala, S., Chowdhury, A., Dhall, A., Goyal, P. (eds) Computer Vision and Image Processing. CVIP 2021. Communications in Computer and Information Science, vol 1567. Springer, Cham*. https://doi.org/10.1007/978-3-031-11346-8_44

[4] S. Sivanandham, D.B. Gunaseelan. (2022). Development of an Ensembled Meta-Deep Learning Model for Semantic Road-Scene Segmentation in an Unstructured Environment. *Applied Sciences*, 12(23), 12214. https://doi.org/10.3390/app122312214

[5] D. Bhattacharya, A. Bhattacharyya, M. Agrebi, A. Roy, P.K. Singh. (2025). DFE-AVD: deep feature ensemble for automatic vehicle detection. *In: Biswas, S.K., Bandyopadhyay, S., Hayashi, Y., Balas, V.E. (eds) Intelligent Computing Systems and Applications. ICICSA 2023. Lecture Notes in Networks and Systems, vol 1307. Springer, Singapore*. https://doi.org/10.1007/978-981-96-3860-4_27

[6] A. Gupta, A. Anpalagan, L. Guan, A.S. Khwaja. (2021). Deep learning for object detection and scene perception in self-driving cars: Survey, challenges, and open issues. *Array*, 10, 100057. https://doi.org/10.1016/j.array.2021.100057

[7] L. P. Osco, J. Marcato Junior, A. P. M. Ramos, L. A. de Castro Jorge, S. N. Fatholahi, J. de Andrade Silva, E. T. Matsubara, H. Pistori, W. N. Gonçalves, and J. Li (2021). A review on deep learning in UAV remote sensing. *International Journal of Applied Earth*

Observation and Geoinformation, 102, 102456. https://doi.org/10.1016/j.jag.2021.102456

[8] S. Kolekar, S. Gite, B. Pradhan, K. Kotecha. (2021). Behavior Prediction of Traffic Actors for Intelligent Vehicle Using Artificial Intelligence Techniques: A Review. *IEEE Access*, 9, 135034-135058. doi: 10.1109/ACCESS.2021.3116303

[9] S. Kuutti, R. Bowden, Y. Jin, P. Barber, S. Fallah. 2020. A survey of deep learning applications to autonomous vehicle control. IEEE Transactions on Intelligent Transportation Systems, 22(2), 712-733. doi: 10.1109/TITS.2019.2962338

[10] Z.-X. Xia, W.-C. Lai, L.-W. Tsao, L.-F. Hsu, C.-C.H. Yu, H.-H. Shuai, W.-H. Cheng. (2021). A Human-Like Traffic Scene Understanding System: A Survey. *IEEE Industrial Electronics Magazine*, 15(1), 6-15, doi: 10.1109/MIE.2020.2970790

[11] S. Mandal, S. Biswas, V. E. Balas, R. N. Shaw and A. Ghosh. (2020). Motion prediction for autonomous vehicles from Lyft dataset using deep learning. 2020 IEEE 5th International Conference on Computing Communication and Automation (ICCCA), Greater Noida, India, 2020, pp. 768-773, doi: 10.1109/ICCCA49541.2020.9250790

[12] H. Soori, D. Khorasani-Zavareh. (2019). Road traffic injuries measures in the Eastern Mediterranean Region: Findings from the Global Status Report on Road Safety–2015. *Journal of Injury and Violence Research*, 11(2), 149-158. doi: 10.5249/jivr.v11i2.1122

[13] Z. Wang, W. Ren, Q. Qiu. (2018). LaneNet: Real-time lane detection networks for autonomous driving. *ArXiv*, abs/1807.01726

[14] P. Deoli, R. Kumar, A. Vierling, K. Berns. (2024). Evaluating the Robustness of Off-Road Autonomous Driving Segmentation against Adversarial Attacks: A Dataset-Centric analysis. *ArXiv*. DOI:10.48550/arXiv.2402.02154

[15] H. Xu, H. He, Y. Zhang, L. Ma, J. Li. (2023). A comparative study of loss functions for road segmentation in remotely sensed road datasets. *International Journal of Applied Earth Observation and Geoinformation*, 116, 103159. https://doi.org/10.1016/j.jag.2022.103159

[16] B. Baheti, S. Innani, S. Gajre, S. Talbar. (2020). Eff-UNet: A Novel Architecture for Semantic Segmentation in Unstructured Environment. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Seattle, WA, USA, 2020, pp. 1473-1481*. doi: 10.1109/CVPRW50498.2020.00187

[17] Y. Yuan, X. Chen, J. Wang. (2020). Object-Contextual Representations for Semantic Segmentation. *In: Vedaldi, A., Bischof, H., Brox, T., Frahm, JM. (eds) Computer Vision – ECCV 2020. ECCV 2020. Lecture Notes in Computer Science, vol 12351. Springer, Cham*. https://doi.org/10.1007/978-3-030-58539-6_11

[18] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. Álvarez, P. Luo. (2021). SegFormer: Simple and efficient design for semantic segmentation with transformers. *ArXiv*, abs/2105.15203.

[19] J. Uhrig, M. Cordts, U. Franke, T. Brox. (2016). Pixel-level Encoding and Depth Layering for Instance-level Semantic Labeling. *In: Rosenhahn, B., Andres, B. (eds) Pattern Recognition. GCPR 2016. Lecture Notes in Computer Science, vol 9796. Springer, Cham*. https://doi.org/10.1007/978-3-319-45886-1_2

[20] V. Badrinarayanan, A. Kendall, R. Cipolla. (2017). SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(12), 2481-2495. doi: 10.1109/TPAMI.2016.2644615

[21] L.-C. Chen, G. Papandreou, F. Schroff, H. Adam. (2017). Rethinking atrous convolution for semantic image segmentation. *ArXiv*, abs/1706.05587.

[22] V. Iglovikov, A.A. Shvets. (2018). TernausNet: U-Net with VGG11 encoder pre-trained on imagenet for image segmentation. *ArXiv*,

abs/1801.05746.

[23] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, H. Adam. (2018). Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. *In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds) Computer Vision – ECCV 2018. ECCV 2018. Lecture Notes in Computer Science, vol 11211. Springer, Cham*. https://doi.org/10.1007/978-3-030-01234-2_49

[24] S. Mi, Q. Bao, Z. Wei, F. Xu, W. Yang. (2021). MBFF-Net: Multi-Branch Feature Fusion Network for Carotid Plaque Segmentation in Ultrasound. *In: de Bruijne, M., et al. Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. MICCAI 2021. Lecture Notes in Computer Science, vol 12905. Springer, Cham*. https://doi.org/10.1007/978-3-030-87240-3_30

[25] M. Fan, S. Lai, J. Huang, X. Wei, Z. Chai, J. Luo, X. Wei. (2021). Rethinking BiSeNet For Real-time Semantic Segmentation. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), USA, 2021, pp. 9711-9720*. doi: 10.1109/CVPR46437.2021.00959

[26] J. Wang, E. Olson. (2016). AprilTag 2: Efficient and robust fiducial detection. *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Daejeon, Korea (South), 2016, pp. 4193-4198*. doi: 10.1109/IROS.2016.7759617

[27] S. Khairnar, S. Gite, K. Kotecha, S.D. Thepade. (2023). Face Liveness Detection Using Artificial Intelligence Techniques: A Systematic Literature Review and Future Directions. *Big Data and Cognitive Computing*, 7(1), 37. https://doi.org/10.3390/bdcc7010037

[28] S. Khade, S. Ahirrao, S. Phansalkar, K. Kotecha, S. Gite, S.D. Thepade. (2021). Iris Liveness Detection for Biometric Authentication: A Systematic Literature Review and Future Directions. *Inventions*, 6(4), 65. https://doi.org/10.3390/inventions6040065

[29] D. Singh, Y.S. Taspinar, R. Kursun, I. Cinar, M. Koklu, I.A. Ozkan, H.-N. Lee. (2022). Classification and Analysis of Pistachio Species with Pre-Trained Deep Learning Models. *Electronics*, 11(7), 981. https://doi.org/10.3390/electronics11070981

[30] Z. Wang, M. Xie, Yi Lin, T. Wu. (2024). A Study on the Effectiveness of Deep Learning Architectures in Style Transfer: A Comparative Analysis of CNN, VGG16, and VGG19. *Proceedings of the 2023 5th International Conference on Big Data Service and Intelligent Computation (BDSIC '23). Association for Computing Machinery, New York, NY, USA*, 82-93. https://doi.org/10.1145/3633624.3633636