Computers, Materials & Continua

Tech Science Press

Doi:10.32604/cmc.2025.064924

ARTICLE

Visual Perception and Adaptive Scene Analysis with Autonomous Panoptic Segmentation

Darthy Rabecka V[1,*], Britto Pari J[1] and Man-Fai Leung[2,*]

[1]School of Electrical and Communication, Department of ECE, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Chennai, 600062, India

[2]School of Computing and Information Science, Department of Science and Engineering, Anglia Ruskin University, Cambridge, CB1 1PT, UK

*Corresponding Authors: Darthy Rabecka V. Email: vtd1112@veltech.edu.in; Man-Fai Leung. Email: man-fai.leung@aru.ac.uk

Received: 27 February 2025; Accepted: 04 July 2025; Published: 29 August 2025

ABSTRACT: Techniques in deep learning have significantly boosted the accuracy and productivity of computer vision segmentation tasks. This article offers an intriguing architecture for semantic, instance, and panoptic segmentation using EfficientNet-B7 and Bidirectional Feature Pyramid Networks (Bi-FPN). When implemented in place of the EfficientNet-B5 backbone, EfficientNet-B7 strengthens the model's feature extraction capabilities and is far more appropriate for real-world applications. By ensuring superior multi-scale feature fusion, Bi-FPN integration enhances the segmentation of complex objects across various urban environments. The design suggested is examined on rigorous datasets, encompassing Cityscapes, Common Objects in Context, KITTI Karlsruhe Institute of Technology and Toyota Technological Institute, and Indian Driving Dataset, which replicate numerous real-world driving conditions. During extensive training, validation, and testing, the model showcases major gains in segmentation accuracy and surpasses state-of-the-art performance in semantic, instance, and panoptic segmentation tasks. Outperforming present methods, the recommended approach generates noteworthy gains in Panoptic Quality: +0.4% on Cityscapes, +0.2% on COCO, +1.7% on KITTI, and +0.4% on IDD. These changes show just how efficient it is in various driving circumstances and datasets. This study emphasizes the potential of EfficientNet-B7 and Bi-FPN to provide dependable, high-precision segmentation in computer vision applications, primarily autonomous driving. The research results suggest that this framework efficiently tackles the constraints of practical situations while delivering a robust solution for high-performance tasks involving segmentation.

KEYWORDS: Panoptic segmentation; multi-scale features; efficient net-B7; Feature Pyramid Network

1 Introduction

The study [1] SAM4UDASS uses semantic-guided mask labeling and fusion algorithms for integrating SAM with unsupervised domain adaptation. It improves the segmentation accuracy of small objects and infrequent classes in intelligent cars. However, it adds computing complexity and depends on SAM's pre-trained masks. SAM and Deep SORT are employed in this study [2] to detect and monitor vehicles in low-resolution, un-calibrated urban surveillance films. It exceeds standard methods like YOLOv8 and SSD in terms of accuracy. However, real-time deployment is restricted by its complexity. Models including Deep Residual Networks (Res Nets) have heightened segmentation accuracy by tackling vanishing gradient obstacles and permitting deeper networks that can capture intricate visual details [3]. The suggested method, which integrates deep learning models, sensor fusion, and human-inspired reasoning, exceeds augmented intelligence-based methods for autonomous driving. In confusing traffic situations, this strategy improves decision-making and overtaking safety. But in sensor-limited or uncertain scenarios, it can be difficult to gather high-quality, real-time data, which is an essential requirement of this approach [4]. Faster R-CNN [5] highlighted the practical application of real-time region proposal networks (RPNs) for faster identification of objects and segmentation, but with a substantial trade-off between speed and accuracy, as its bottom-up nature allows performance to decrease for small things. Autonomous driving is enabled by deep reinforcement learning [6], which reduces the need for expensive sensors by relying on visual information gathered from cameras. This holistic approach not only lowers hardware costs but also simplifies system design. However, it requires extensive training and often struggles to generalize across varying conditions. Significant advancements in this field were made by Swin Transformers, who introduced hierarchical designs with moving windows that made it possible to create segmentation models that were both scalable and effective, and which demonstrated excellent performance across a range of applications [7] demanding an excessive amount of training time and a lot of training data. To accomplish state-of-the-art results, Max-Deep Lab fully combined segmentation and mask prediction tasks through the use of end-to-end panoptic segmentation leveraging mask transformers [8] focuses more on evaluating model architecture than on practical implementation.

Pre-trained transformer models revealed better generalization to various kinds of unknown datasets and strengthened the robustness of panoptic algorithmic segmentation [9], but they have slow inference and high GPU use of memory. Real-world domains where accurate scene perception is crucial, such as autonomous driving, can benefit from panoptic segmentation. Benchmarks like the Cityscapes dataset [10] have facilitated the development of robust segmentation methods for dynamic scenarios. It needs multiple sensor data to be blended and coordinated. Among the innovative ways that scholars have developed to improve current frameworks are semantic segmentation heads, which seamlessly communicate with instance segmentation components to generate more coherent segmentation outputs [11], but they have high inference-related memory usage. Feature Pyramid Networks (FPNs) have been essential for consistent detection and segmentation throughout an array of object sizes and scales in multi-scale feature aggregation [12]. It's difficult to train such multitask models. This approach [13] integrates multi-agent collaborative bird's-eye-view segmentation with domain generalization techniques to address scene generalization. This technique facilitates coordination and robustness in an unknown environment. Yet, its use in real-time is challenging due to its high computational power and requirement for enormous data sets.

Panoptic-Deep Lab created a proposal-free approach that balanced accuracy and efficiency by regressing instance centers and offsets, and Feature Pyramid Networks (FPNs) successfully combined dense bounding boxes into instance masks [14]. However, this method has limited adjustability for complex multi-class instances. A hybrid deep learning system [15] that combines Faster R-CNN, SSD, and YOLO has been proposed to enhance lane and object recognition for autonomous cars. This model leverages the unique strengths of each detector to improve accuracy and resilience across various traffic situations. However, the computational complexity of integrating multiple architectures can pose challenges for real-time deployment on low-power platforms. By employing a novel efficient panoptic segmentation technique and replacing the EfficientNet-B5 backbone with EfficientNet-B7, the primary objective of this study is to enhance performance for instance segmentation, semantic segmentation, and panoptic segmentation tasks [16]. It handles circumstances having intricate object associations. A practical and efficient replacement for classic impact load recognition methods is offered in this research [17], signifying an important development in structural health monitoring. The incorporation of proficient visual sensing technology opens the door to high-precision, real-time, non-contact structural integrity assessment. Lighting and surrounding variables have an impact on accuracy.

The study [18] introduces an automated vision-based strategy for detecting bridge displacements on multiple planes, enabling precise measurements without physical contact via extensive image processing techniques. The effectiveness of structural monitoring is boosted by this imaginative approach, which also ensures enhanced security and cost-effectiveness, but it has issues with standpoint sensitivity and camera calibration. This work describes an improved technique [19] for intelligent visual perception-based moving force identity, which permits contact-free and real-time analysis. This technique involves the use of complex computer vision techniques to ensure excellent identification of dynamic loads on structures. It has significant effects on improving monitoring accuracy and structural safety; however, its real-time dynamic scene interpretation introduces complications.

Autonomous systems need accurate visual attention in order to properly navigate challenging configurations. The precision, performance, and adaptability of current solutions are often compromised. This work overcomes these issues holistically through the implementation of a scalable and robust construction that incorporates panoptic segmentation, Bi-FPN, and EfficientNet-B7. The EfficientNet-B7 backbone proposed here outperforms existing approaches on several benchmark datasets, including Cityscapes, COCO, KITTI, and IDD. Incorporating efficient panoptic segmentation components further improves its ability to handle a range of urban and real-world environments with remarkable precision while maintaining lower processing requirements. The primary contributions to this work have been outlined as follows:

- Backbone EfficientNet-B7: Using EfficientNet-B7 as the backbone as opposed to EfficientNet-B5, the study aims to improve feature extraction and the performance of semantic, instance, and panoptic segmentation tasks.
- Multi-Scale Feature Fusion with Bi-FPN: Bi-FPN improves the segmentation of complex objects in urban and real-world environments by facilitating better multi-scale feature fusion, which raises the model's overall accuracy and efficiency.
- Customized Efficient Panoptic Segmentation Technique: A novel approach to panoptic segmentation is introduced, with the goal of optimizing the segmentation process and achieving high precision while reducing processing demands for real-time applications.

On benchmark datasets such as Cityscapes, COCO, KITTI, and IDD, the proposed method achieves state-of-the-art performance, surpassing existing segmentation methods in terms of accuracy and efficiency.

2 Related Works

A solid, simple baseline with quick bottom-up segmentation is provided by Panoptic-Deep Lab [5]. Its incapacity to deal with small object details, however, limits the segmentation quality in scenes with a lot of people. Even if it works well, real-time applications need a lot of processing power. To reduce the reliance on additional sensors like LiDAR, a deep reinforcement learning system [6] was developed to enable autonomous navigation using only visual input. This approach allows for end-to-end policy learning from initial images, providing a small and cost-effective solution. However, it has limitations in adapting to rapidly changing driving situations and requires lengthy training cycles. Attention Mask with Mask Transformer [11] uses attention strategies to segment images successfully, enhancing generalization across segmentation tasks. It handles huge images slowly and requires a lot of memory. The multi-task network [12] simplifies feature sharing by correctly handling panoptic segmentation for automatic driving. However, its capacity for generalization across datasets is restricted since it relies on high-quality annotations. Deployment on low-power devices is restricted by its transformer-based approach's high computational need for resources. Enhancing feature fusion through the use of soft attention processes, the Fast Panoptic Segmentation with Soft Attention Embeddings [20] technique enables greater differentiation of overlapping items and increased segmentation accuracy. Since soft attention embeddings are costly to compute, this system has difficulties processing in real-time in complicated situations. To ensure that, for instance, semantic regions are segmented, a method that accurately locates object centers is presented in the Center-Guided Transformer for Panoptic Segmentation [21]. Its slower inference time renders it less suited for latency-critical demands than lightweight CNN-based alternatives. Deeper Lab [22] greatly improves efficiency by delivering outstanding segmentation accuracy using a single-shot method. Its reliance on more complex structures, however, enhances processing costs and reduces its practicality for real-time applications. An adapted instance selection conduct is introduced by Adapt IS [23], enhancing segmentation accuracy for objects with various scales. Its region proposal-based method, however, extends the inference time and decreases its efficiency for real-time applications. By the creation of effective network architecture [24], this method accelerates panoptic segmentation while maintaining accuracy and reducing technology overhead. Nevertheless, since feature aggregation is simplified, it performs inferiorly in cluttered environments. In autonomous driving, deep learning models [25] are commonly used for tasks such as object detection, semantic segmentation, and decision-making. These models offer high accuracy and enable end-to-end learning from camera inputs, enhancing perception and facilitating real-time control. However, they demand significant computational resources and rely on annotated datasets. Additionally, their black-box nature poses challenges for debugging and interpretation. By utilizing comprehensive data augmentation, hyper parameter tuning, and domain adaptation, AI training optimization strategies [26] enhance model performance. These methods reduce training time while improving the model's ability to generalize across various driving environments. However, due to unseen domain shifts, these advantages do not always translate into significant real-world performance improvements and still depend on the availability of diverse, high-quality data. This method [27] utilizes pixel consensus voting to enhance the accuracy of panoptic segmentation, particularly for ambiguous regions. However, it is less suitable for real-time applications due to the added analyzing complexity. Inference time decreases [28] while competitive performance is maintained with a single-shot panoptic segmentation model. However, in contrast with multi-stage strategies, it compromises some segmentation granularity.

The recommended strategy employs risk-aware planning and probabilistic modeling to guarantee safe actions in unexpected multi-vehicle scenarios. This method [29] optimizes safety in volatile situations. Real-time responses are slowed down by their main limitation, which is the increased computing strain. Feature Panoptic Pyramid networks [30] improve instance and semantic segmentation by using multi-scale feature representations. However, the computing cost is increased by the method's complexity. There have been creative attempts to enhance panoptic segmentation performance, and this area of study is still being researched.

3 Methods

Ideal for panoptic segmentation challenges, the suggested autonomous PS architecture is made to generate precise as effective segmentation results. Building on the Efficient Net B7 design, it has a shared backbone and a new Bi-FPN that allows information to flow in both directions for improved feature representation. Parallel heads are incorporated into the design, for instance, and semantic segmentation. The semantic segmentation head uses a special design that is suited to the task, whereas the modified Mask R-CNN topology serves as the foundation for the instance segmentation head. The outputs of each head are extracted independently, and the final panoptic segmentation output is produced by fusing them in the panoptic fusion module. High accuracy and efficiency are ensured by this architecture, which is optimized for autonomous systems and substitutes Efficient Net B7 for the previously utilized Efficient Net B5 [16] while keeping the remaining components.

3.1 Bi-FPN with Efficient Net B7 Architecture as a Backbone

Despite having fewer parameters and FLOPs, Efficient Net can perform significantly better than other networks in classification assessments. To smoothly and frequently scale the network's width, depth, and resolution, it uses compound scaling. Therefore, we improve on this scaled layout, which is additionally referred to as the EfficientNet-B7 model, with a scaled architecture coefficient of 2.0, 2.5, and 600. Other Efficient Net modifications can be rapidly chosen as this, depending on computational limitations and available resources. First, we turn off the Squeeze-and-Excitation (SE) connections and the classification head in order to modify EfficientNet-B7 to meet our goal. SE connections focus on contextual elements more than feature localization, which is excellent for classification but not so good for segmentation tasks; nevertheless, they strengthen interdependencies between channels. We exclude SE connections into our backbone since both factors are equally important in segmentation. We next employ synchronized *In Situ* Activated Batch Normalization (iABN Sync) to replace all of the Batch Normalization (Batch Norm) layers. Thus synchronizes batch statistics across GPUs, enhancing gradient estimation in multi-GPU training while optimizing memory use. As shown in Fig. 1, nine blocks make up our EfficientNet-B7.

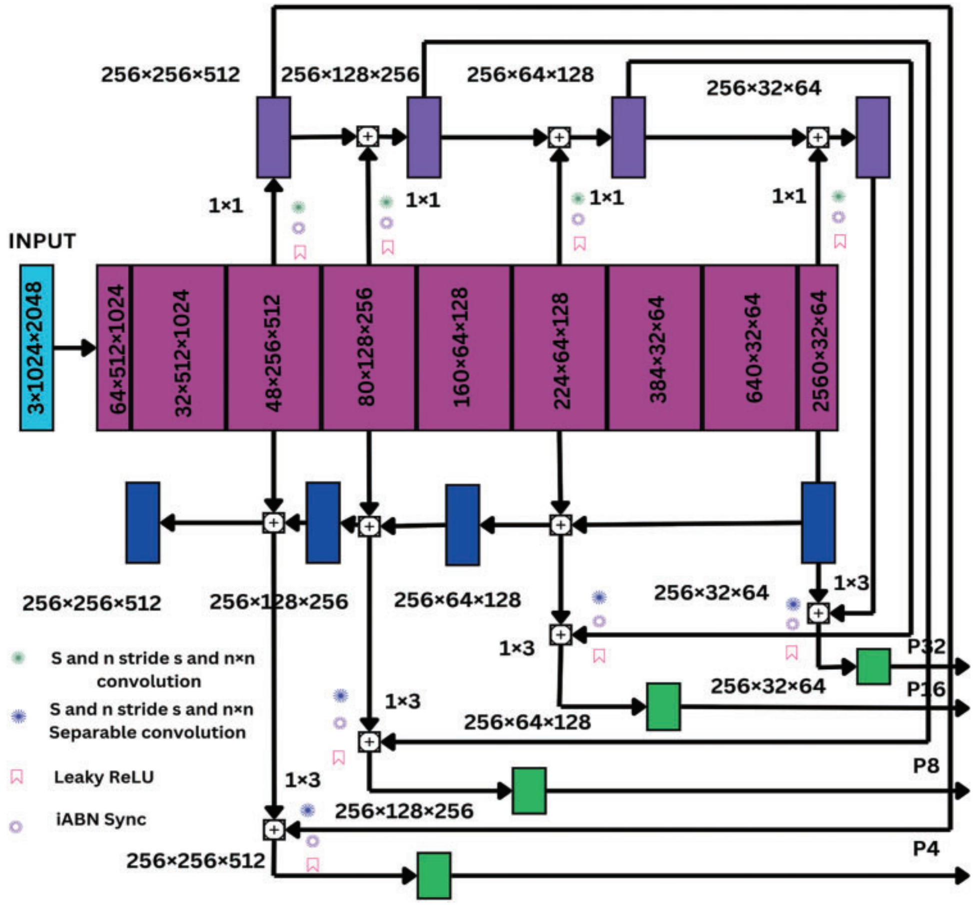


Figure 1: An example of our suggested effective B7 architecture used as the backbone of the Bi-Feature Pyramid Network is shown. The Bi-FPN ensures a bi-directional flow of information

The ×4, ×8, ×16, and ×32 downsampling factors, correspondingly, are represented by the outputs of blocks 2, 3, 5, and 9. Our suggested Feature Pyramid Network (FPN) in two directions receives these down-sampled outputs. Conventional FPNs in panoptic segmentation use a top-down method to collect multi-scale characteristics. First, a 1 × 1 convolution is used to set the number of channels in encoder outputs to a predetermined worth, typically 256. Following that, features with lower resolution are up-sampled and merged with those with better resolution. To combine ×16 resolution encoder outputs, for example, ×32 resolution features must first be decreased to ×16 resolutions. Each scale then undergoes a 3 × 3 convolution in order to enhance feature fusion, leading to P4, P8, P16, and P32 outputs. However, effective multi-scale feature fusion is constrained by this unidirectional feature aggregation. We provide a Bi-FPN that combines top-down and bottom-up information flow in order to overcome this constraint. Two parallel branches make up our Bi-FPN, and each one uses 256 output filters at each scale in a 1 × 1 convolution for channel reduction.

After capturing higher-resolution features from left to right, the bottom-up branch down blends them with equivalent lower-resolution encoder outputs, whereas the top-down branch utilizes the conventional right-to-left FPN aggregation. Features with ×4 resolution, for example, are enlarged to ×8 resolutions and appended to encoder outputs with ×8 resolution. Finally, the Outputs for P4, P8, P16, and P32 are obtained by summing the top-down and bottom-up outputs at each resolution before passing them through a 3 × 3 depth-wise separable convolution with 256 output channels. For lowering the expenses of parameters, we adopt depth-wise separable convolutions rather than normal convolutions. Our model translates from EfficientNet-B5 to EfficientNet-B7 and enhances multi-scale fusion, feature extraction, and segmentation performance through the use of EfficientNet-B7's most effective scaling properties.

The bottom-up branch samples higher-resolution features from left to right before merging them with corresponding lower-resolution encoder outputs, whereas the top-down branch utilizes the conventional right-to-left FPN aggregation. Features with ×4 resolution, for example, are enlarged to ×8 resolutions and appended to encoder outputs with ×8 resolution. Finally, the P4, P8, P16, and P32 outputs are obtained by summing the top-down and bottom-up outputs at each resolution before passing them through a 256-channel 3 × 3 depth-wise separable convolution. For lowering the expenses of parameters, we adopt depth-wise separable convolutions rather than normal convolutions. Our model translates from EfficientNet-B5 to EfficientNet-B7 and enhances multi-scale fusion, feature extraction, and segmentation performance through the use of EfficientNet-B7's most effective scaling properties.

3.2 Integration of Efficient Net B7 and Bi-FPN Architecture with Modified Efficient PS Branch

EfficientNet-B7 operates as the backbone for this architecture, in addition to Bi-FPN for multi-scale feature fusion and an Efficient Panoptic Segmentation Branch that combines instance and semantic segmentation to create one output for real-time applications such as autonomous driving.

As shown in Fig. 2, the proposed structure enables multi-scale feature representation through the addition of Bi-FPN on top of the EfficientNet-B7 backbone. After processing the input image through the backbone, four spatial levels of features, P4, P8, P16, and P32 are acquired. Using 1 × 1 convolutions, these features, which fundamentally fluctuate in resolution and channel depth, are channel-aligned to ensure uniform dimensionality before fusion. Enabling accurate feature combining across scales requires this alignment. By employing both top-down and bottom-up avenues to improve features, Bi-FPN proposes a bidirectional fusion method. The top-down way uses learnable weighted addition to up sample and integrates high-level semantic features from deeper layers (like P32) with lower-level features (like P16, P8, and P4). In the bottom-up manipulation, the semantically rich high-level features are paired back up with spatially detailed low-level features. The model has can bring out task-relevant features at each stage of this iterative feature fusion. To decrease the computational load while maintaining accuracy, depth-wise separable convolutions are employed after each fusion step. The network can adaptively prioritize features from different categories as the fusion nodes are regulated by learnable attention weights. For enhanced generalization and stability, synchronized batch normalization (iABN Sync) and leaky ReLU activations are used. By efficiently collecting both fine and coarse facts, this Bi-FPN design provides an excellent framework for the panoptic branch's instance and semantic segmentation heads.

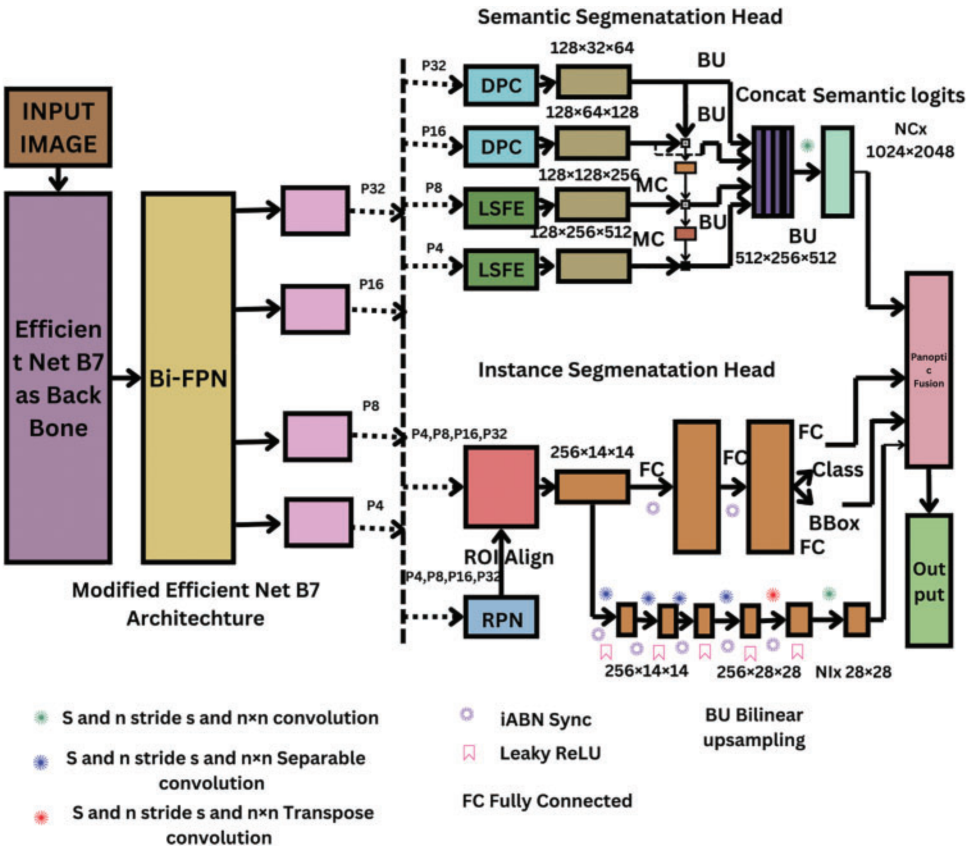


Figure 2: A demonstration of our intended Bi-FPN and Efficient Net B7 created with a powerful PS Branch [16]

3.3 Semantic Segmentation Head

The three primary elements of our hypothesized semantic segmentation head, which is based on existing panoptic segmentation methods, are each supposed to address an essential requirement. The network is required to; first of all, efficiently capture fine-grained characteristics at a vast scale. The two 3 × 3 depth-wise separable convolutions with 128 output filters each make up the Large Scale Feature Extractor (LSFE) module, which is integrated to do this. Leaky ReLU activation function and in-place activated batch normalization (iABN sync) are executed following these convolutions. The next convolution further refines deeper feature representations, though. The filter count is reduced to 128 by the first 3 × 3 depth-wise separable convolution. As apparent in Fig. 3, we make use of an altered DPC module within our semantic head. Leaky Re LU is employed for Re LU activation and iABN sync is deployed to substitute batch normalization layers in the original DPC architecture. A 256-channel depth-wise separable convolution with a dilation rate of (1, 6) in 3 × 3 comprises the DPC module, which draws on the Efficient Panoptic Segmentation [17] (Efficient PS) algorithms recently in existence. It spreads into five parallel branches. The dilation rate of the 256 outputs of the 3 × 3 dilated depth-wise separable convolution in the first branch is (1, 1). The second branch has a 3 × 3 dilated depth-wise separable convolution with 256 outputs and a dilation rate of (6, 21).

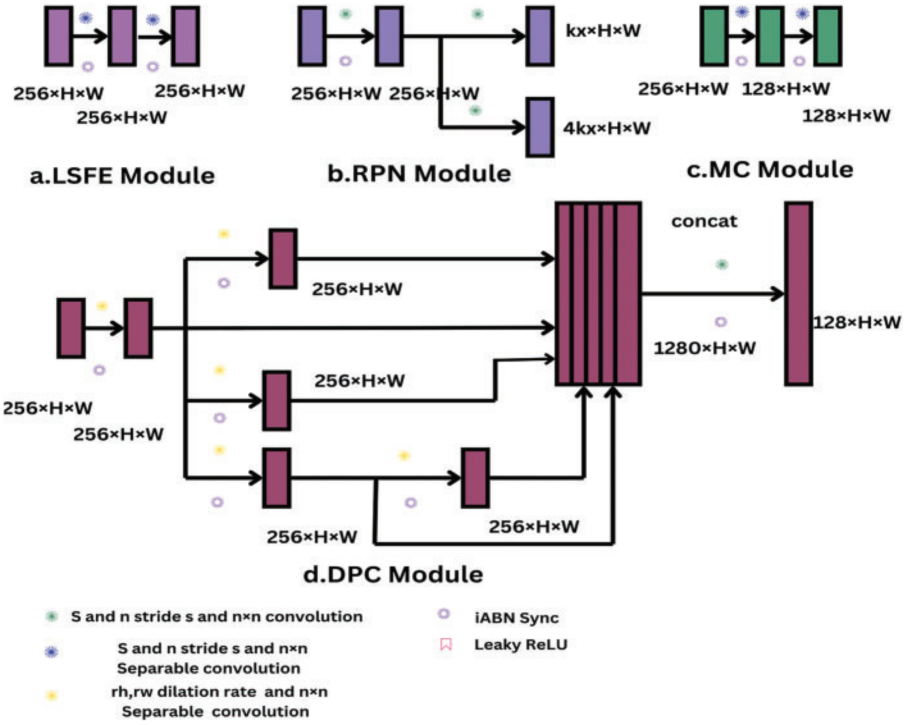


Figure 3: Our proposed architecture utilizes existing efficient panoptic segmentation, which involves various topologies of semantic and instance Heads

A 256-output 3 × 3 dilated depth-wise separable convolution with a dilation rate of (18, 15) is found in the third branch. With 256 outputs and a dilation rate of (6, 3), the fourth branch takes the outcome of the third branch (dilation rate 18, 15) and runs it via an additional 3 × 3 dilated depth-wise separable convolution. Joining the outputs from each branch builds a 1280-channel tensor, which is then transmitted through the DPC module's final output, produced via a 1 × 1 convolution with 256 output channels. A leaky ReLU activation function with iABN sync follows each convolution in the DPC module. The capacity to minimize the disparity between max-scale and min-scale features during feature execution is the triad and last requisite for the semantic head aggregate. We achieve this by employing our Mismatch Correction Module (MC), which links tiny-scale attributes to huge-scale features. The feature maps are up-sampled by a factor of two utilizing a bilinear up-sampling layer after iABN sync with Leaky ReLU and cascaded 3 × 3 depth-wise separable convolutions with 128 output channels. Fig. 3a,c,d shows the structure of these primary parts. This makes up our semantic head in this proposed architecture. The four different scaled outputs of our Bi-FPN, P4, P8, P16, and P32, are the inputs to our semantic head. The large-scale inlets, P8 and P4, with down-sampling elements of ×8 and ×4, are each submitted via two equivalent LSFE modules, while the small-measure inputs, P32 and P16, with downgrading values of ×32 and ×16, are fed into two similar DPC components. The four independent scaled outputs of our Bi-FPN (P4, P8, P16, and P32) serve as the inputs of our semantic head.

The outputs of these associated LSFE and DPC modules are then up-sampled to ×4 scales and supplemented by feature alignment connections. A 512-channel tensor is generated by connecting these up-sampled feature maps, and it is then fed into N-stuff output filters combined with 1 × 1 convolutions. The semantic logits are subsequently obtained with the same resolution as the supplied image by up-sampling the tensor by a factor of four and passing it through a soft max layer. As seen in Fig. 3, these outputs are connected element-wise adding via connectors for feature alignment between the DPC and LSFE modules. We present our MC modules in the connections between the two LSFE connections and the second DPC and LSFE. To enhance the refinement of object limitations, these correlation associations combine dependent data from small-scale properties with distinguishing large-scale properties. For training, we employ the weighted per-pixel log loss, which is provided by

$$L_{pp}\left(\Theta\right) = -\sum_{ij} w_{ij}(p_{ij}^{*}) \log p_{ij} \tag{1}$$

The ground truth for an image is represented by p_{ij}^{*} and the expected likelihood for the pixel (i, j) being assigned class c ∈ p is represented by $P_{i,j}$. $w_{i,j} = 0$ otherwise, and $w_{i,j} = 4/\mathrm{WH}$ if pixel (i, j) falls within 25% of the worst forecast. W and H denote the supplied input image's width and height, respectively.

The following provides the overall semantic head loss:

$$L_{semantic}\left(\Theta\right) = \frac{\mathbf{1}}{n} \sum L_{pp}, \tag{2}$$

where n is the batch size.

Two convolution layers and two In-Place ABN layers are used in the initial LSFE module to extract features. Depth-wise separable convolutions were used to gather depth-wise and spatial correlations in the first convolution layer (conv_1). Abn_1 is followed by normalization and non-linearity. Abn_2 incorporates normalization and non-linearity, while the second convolution layer, conv_2, further processes the feature maps. The three additional layers in the new LSFE module—convolution layer conv_2, convolution layer conv_3, and its corresponding In-Place ABN layer, abn_3—increase the model's ability to recognize complex patterns and higher-level semantic representations. The revised architecture has better learning hierarchy features that are expected to boost performance on demanding tasks.

3.4 Instance Segmentation Head

With some alterations, this recommended Autonomous PS model's instance segmentation head possesses a topology that is similar to an existing framework. Leaky ReLU, iABN sync, and depth-wise separable convolution are applied in place of ReLU activations, standard convolutions, and batch normalization layers. In order to save 2.09 million parameters over the previous framework, depth-wise separable convolutions are included throughout the process of execution. The two-stage structure of the recommended structure remains intact. Creating object proposals is the domain of the Region Proposal Network (RPN), making it the first step. During the improvement of these ideas, the second stage carries out mask prediction, bounding box regression, and classification. For optimization, there are five loss functions included. In RPN, the objectness score loss L_{os} is a log loss for objectness classification, with preset IoU thresholds identifying positive and negative matches:

$$L_{os}\left(\Theta\right) = -\frac{1}{|N_s|} \sum_{(P_{os}^{*}, P_{os}) \in Ns} P_{os}^{*} \cdot \log P_{os} + (1 - P_{os}^{*}) \cdot \log(1 - P_{os}), \tag{3}$$

where P_{os}, the output of the objectness score branch of RPN and P^*_{os} denotes the ground truth label, which is 0 if the anchor is negative and 1 if the anchor is positive. Positive and negative matches are evaluated using the preset thresholds T H and T L. The object proposal loss is computed using the smooth L_1 norm and is a regression loss that is exclusively applied to positive matches.

$$L_{op}(\Theta) = -\frac{1}{|N_s|} \sum_{(t^*,t)\in Np} \sum_{(i^*,i)\in(t^*,t)} L_1(i^*, i), \tag{4}$$

The parameterizations of the ground truth and predicted bounding boxes are provided by where N_s is the subset of positive matches N_P.

$$t_x = \frac{x - x_a}{w_a}, t_y = \frac{y - y_a}{h_a}, t_w = \log\frac{w}{w_a}, t_h = \log\frac{h}{h_a} \tag{5}$$

$$t^*_x = \frac{x^* - x_a}{w_a}, t^*_y = \frac{y^* - y_a}{h_a}, t^*_w = \log\frac{w^*}{w_a}, t^*_h = \log\frac{h^*}{h_a} \tag{6}$$

where the anchor box parameters are x_a, y_a, w_a, and h_a. Adding a cross-entropy loss to a randomly picked group of positive and negative matches is the classification loss L_{cls}:

$$L_{cls}(\Theta) = -\frac{1}{|k_s|} \sum\nolimits_{c=1}^{N\cdot thing'+1} Y^*_{o,c} \cdot \log Y_{o,c,}\text{for } (Y^*, Y) \in K_s, \tag{7}$$

$|k_s|$ is the number of randomly chosen positive and negative matches. Y represents the classification branch's output, o is the observed class, Y_o is the one-hot encoded ground truth label, and c is the object o correct categorization.

$$L_{bbx}(\Theta) = -\frac{1}{|k_s|} \sum_{(T^*,T)\in k_p} \sum_{(i^*,i)\in(T^*,T)} L_1(i^*, i) \quad , \tag{8}$$

where T^* is the ground truth parameterization, T is the predicted bounding box parameterization, K_p is the set of positive matches, and T^* is the smooth L^* norm.

$$L_{mask}(\Theta) = -\frac{1}{|k_s|} \sum_{(P^*,P)\in Ks} L_P(P^*, P), \tag{9}$$

where $L_P(P^*, P)$ is denoted as,

$$L_P(P^*, P) = -\frac{1}{|T_P|} \sum_{(i,j)\in T_p} P^*_{i,j} \cdot lo + (1 - P^*_{i,j}) \cdot \log\left(1 - P_{i,j}\right), \tag{10}$$

where the loss function L_p compares the anticipated mask P with the ground truth mask P^*. This loss is defined using just positive samples. The total instance segmentation loss is determined by adding the equally weighted losses for objectness score, classification, object proposal, bounding box, and mask segmentation:

$$L_{instance} = L_{os} + L_{op} + L_{cls} + L_{bbx} + L_{mask} \tag{11}$$

The gradients for L_{mask}, L_{cls}, and L_{bbx} are calculated only via the grid backbone and not via the region proposal network, much like in Mask R-CNN. The enhanced B-box Network enhances feature extraction before the final two fully connected layers, which predict bounding box regression and class scores. A third fully connected layer (third-fc) and an additional In-Place ABN layer (abn_3) have been added to improve the expressiveness and learning capability of the network. As part of the Mask Network modifications, four sets of Depth-Wise Separable Conv and In-Place ABN layers are introduced before the de-convolution and final convolution layers. These improvements increase both performance and feature extraction. The output of the Mask Network now includes an extra background channel and pixel-wise mask predictions for each instance. Enhancing bounding box and instance mask feature extraction and prediction accuracy is the main objective of the modifications.

3.5 Panoptic Fusion Module

To generate the PS outlet, the method integrates predictions from the instance segmentation, semantic segmentation, and panoptic fusion heads using Bi-FPN with EfficientNetB7 as the backbone. The instance segmentation head generates bounding boxes, mask logits, class predictions, and confidence ratings. To filter these occurrences, resize and zero-pad mask logits, sort by confidence, and eliminate cases with low confidence ratings. In the second stage, overlapping instances are identified by thresholding mask logits at the sigmoid value 0.5. The instance with the highest confidence is retained. The semantic segmentation head generates semantic logits, and for each instance, the corresponding class channel is selected and masked inside the bounding box. The PS fusion module then combines the semantic and instance logits to produce fused mask logits utilizing Hadamard product computation of the sigmoid of mask logits from both heads and their respective logits. The fused mask logits can be expressed as follows:

$$FL = \left(\sigma\left(ML_A\right) + \sigma\left(ML_B\right)\right) \circ \left(ML_A + ML_B\right) \tag{12}$$

In this case, $\sigma(.)$ represents the sigmoid function, $\circ$ indicates the Hadamard product, represents the instance segmentation mask logit, and represents the semantic segmentation mask logit. These fused logits are concatenated with the 'stuff' logits from the semantic head to produce intermediate panoptic logits after the arg max process to get intercede ps prediction. In the last panoptic segmentation output, a 0-filled canvas is filled with the 'stuff' class predictions from the semantic head and the 'thing' prediction from the instance head. Small area classes are excluded based on a predefined minimum stuff area. The fusion technique adaptively amplifies or attenuates the final instance score based on the agreement or disagreement between the logits, resulting in high-quality panoptic segmentation outputs. Because it uses Bi-FPN with Efficientnet B7 to extract deep features and incorporates segmentation predictions for efficient handling of both 'thing' and 'stuff' classes, it is ideal for applications that require accurate panoptic segmentation. The process is depicted in Fig. 4.

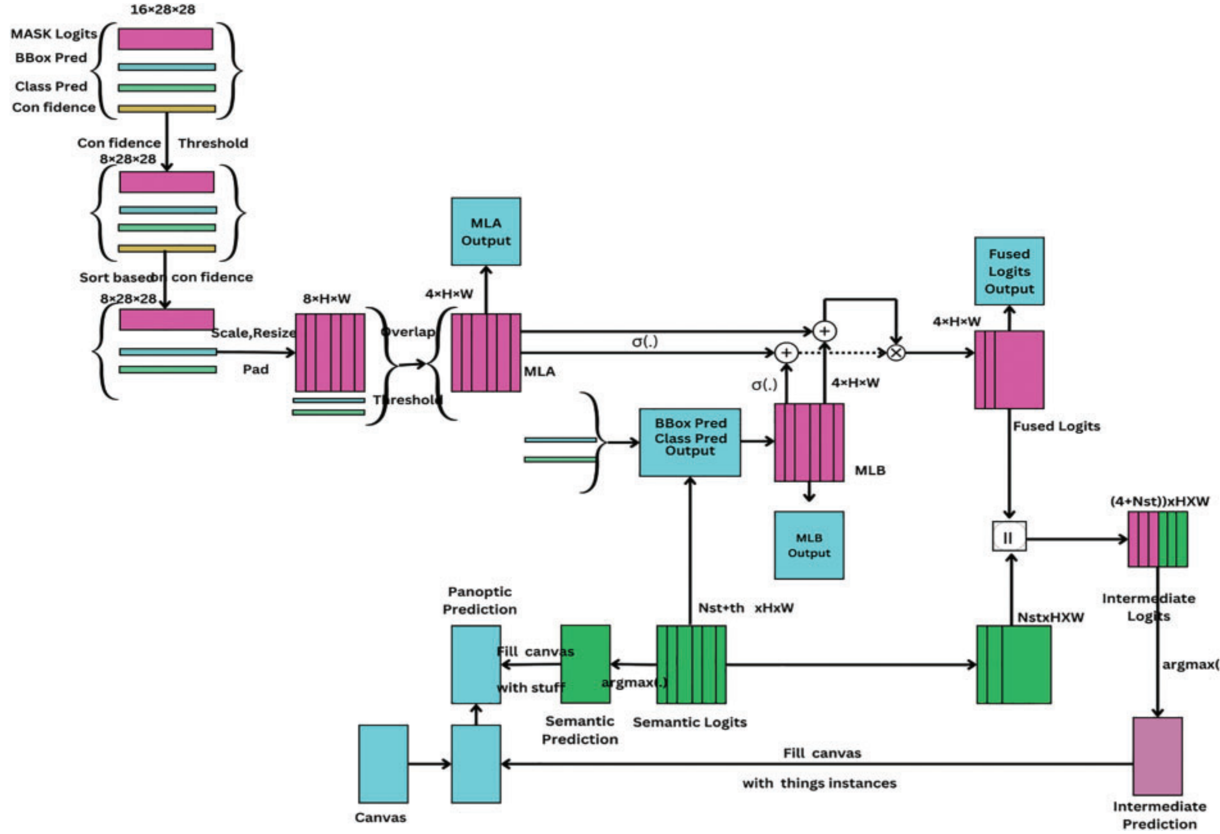


Figure 4: This module combines the mask logits MLA and MLB in the manner described below, σMLA + σMLB $\circ$ (MLA + MLB), where $\circ$ indicates the Hadamard product, $\sigma(.)$ represents the sigmoid function, and MLB is generated by the function f*. The f* function zeroes off the score of a certain class channel in the semantic logits outside of its corresponding bounding box. It's important to keep in mind that the numbers used, such as 4 instances and 16 initial mask logits, are chosen for simplicity even if the true values are far larger

4 Result and Analysis

The dataset we used for our investigations is briefly described in this section, and then the standard metrics we used for evaluations are introduced. After that, the training process is explained. We then present thorough quantitative comparisons and ablation studies of the recommended architectural components. Finally, a visualization of the dataset's panoptic segmentation evaluation is presented.

4.1 Experiment Datasets and Evaluation Metrics

Utilizing EfficientNet-B7 as the foundation, we develop EfficientNet-B7 with Bi-FPN to execute instance, semantic, and panoptic segmentation, building on methods from existing successful panoptic segmentation frameworks. A computer with an Intel Xeon @ 2.20 GHz CPU and NVIDIA TITAN X GPUs is used to train the model using PyTorch.

4.1.1 Cityscapes

The Cityscapes collection, which focuses on urban street scenes, makes driving circumstances easier to understand. It is immensely diverse, with footage from over 50 locations throughout Europe that were filmed in spring, summer, and fall. Dynamic objects, such as cyclists and pedestrians, are often grouped or partially obscured, add to the complexity of the information and complicate panoptic segmentation, especially for the "thing" class. It provides pixel-level annotations for 19 object types, including 11 "stuff" classes and 8 instance-specific "thing" classes. The dataset contains of 5000 highly annotated and 20,000 coarsely annotated images captured at 2048 × 1024 pixels resolution. Of the painstakingly annotated images, 2975 are used for training, 500 for validation, and 1525 for testing. The study's data are publicly accessible in Cityscapes at [20].

4.1.2 COCO Dataset

The COCO dataset comprises over 200,000 images with more than 1.5 million object instances across 80 categories. It provides pixel-level segmentation annotations for the "stuff" and "thing" classes, as well as key point annotations for human posture estimates. The dataset is divided into training, validation, and test sets, with 10,000 images designated for training, 5000 for validation, and 41,000 for testing. The images vary in resolution and depict real-life scenes with extensive backgrounds.

4.1.3 KITTI

The KITTI vision benchmark set provides ground truth for problems involving semantic segmentation, depth prediction, odometry, picture flow estimate, and optical flow estimation. However, there are no signs of panoptic segmentation. To address this, we used the existing Efficient Panoptic Segmentation framework to incorporate panoptic annotations into the KITTI dataset for interpreting urban scenes. With annotations for 8 "thing" classes and 11 "stuff" classes, the dataset consists of 1055 photos (855 for training and 200 for validation), based on the Cityscapes distribution. The annotations were created by combining hand-drawn instance masks with semantic annotations from community-driven KITTI extensions. Both urban and residential settings are captured in the 1280 × 384-pixel quality images. The study's data are publicly available in KITTI. Data from the study are available to the public in the KITTI dataset [16]. This dataset enhances multi-task learning and panoptic segmentation for urban environments.

4.1.4 IDD

The Indian Driving Dataset (IDD) was developed to understand scenes in unstructured scenarios. Unlike other urban datasets, IDD includes scenes with poorly defined infrastructures, such as lanes and walkways. In contrast to previous datasets, it has fewer categories for traffic participants and more "thing" instances. The dataset consists of 10,003 images with 1920 × 1080 or 720 × 1280 pixel size; 6993 of them were utilized for training, 2029 for testing, and 981 for validation. Two Indian cities and their environs served as the locations for the photos. Out of the 26 classes in level 3, "stuff" classes are 17, and 9 are "thing" classes. The results are displayed for this level. There are four tiers of hierarchy for the annotations. The data from the study are publicly available in the IDD dataset [16].

For evaluation, we employ the traditional Panoptic Quality (PQ) metric [16], which is computed as follows:

$$PQ = \frac{\sum_{(p,g)\in TP} IoU\left(p, g\right)}{|TP| + \frac{1}{2}|FP| + \frac{1}{2}|FN|} \tag{13}$$

where *TP*, *FP*, *FN*, and *IoU* represent intersection-over-union, false positives, false negatives, and true positives, respectively. Here is how IoU is defined: $TP/(TP + FP + FN) = IoU$. Additionally, the following metrics are shown for Recognition Quality (RQ) and Segmentation Quality (SQ):

$$SQ = \frac{\sum_{(p,g)\in TP} IoU(p, g)}{|TP|} \tag{14}$$

We report PQ, SQ, and RQ for all classes, as well as metrics for 'stuff' classes (PQ^{st}, SQ^{St}, RQ^{St}) and 'thing' classes (PQ^{Th}, SQ^{Th}, RQ^{Th}) by benchmarking criteria for panoptic segmentation. To be thorough, we additionally provide FLOPs for comparison, average precision (AP), mean intersection-over-union (mIoU) for the "stuff" and "thing" classes, and inference time.

$$RQ = \frac{|TP|}{|TP| + \frac{1}{2}|FP| + \frac{1}{2}|FN|} \tag{15}$$

4.2 Training Protocol

Input image clips in various resolutions—1024 × 2048, 1024 × 1024, 384 × 1280, and 720 × 1280 pixels—are deployed to train the suggested layout. Only a handful of data augmentation techniques are used for training across benchmark datasets, notably Cityscapes, COCO, KITTI, and IDD, to preserve dataset consistency and integrity. These consist of random cropping from the original image resolution, horizontal flipping, and multi-scale resizing with scale information between 0.5 and 2.0. These modifications intend to avoid the creation of inaccurate distortions and retain the reality of driving scenarios. To ensure a fair comparison with other latest methods trained on the same datasets, no additional augmentations—such as color jittering, Gaussian blur, or noise injection are applied. EfficientNet-B7 to initialize the model backbone, as it is pre-trained on the ImageNet dataset. This transfer learning strategy enhances generalization and convergence, particularly for datasets with limited resources. Batch Normalization (BN) strengthens the learning dynamics over all layers. Biases are set to zero, and weight initialization occurs using Xavier initialization. As the activation function for non-backbone layers, Leaky ReLU with a negative slope of 0.01 is utilized. Swish activation is implemented in the backbone to promote a smoother gradient flow. A Bidirectional Feature Pyramid Network (Bi-FPN) is integrated into the model for efficient multi-scale feature fusion. It is also beneficial for segmenting objects at different scales, while this structure permits the network to include both low-level and high-level semantic attributes. We set up the thresholds for the instance segmentation head in the following way:

- TH (high threshold) = 0.7;
- TL (low threshold) = 0.3;
- TN (non-maximum suppression threshold) = 0.5.

Further hyper parameters connected with segmentation include:

- Minimum stuff area (mins) = 2048 pixels;
- Overlap threshold (ot) = 0.5;
- Confidence threshold (ct) = 0.5.

Stochastic Gradient Descent (SGD) with a momentum of 0.9 is applied to the train. The learning rate is modified for every data set to adjust for variability in size, complexity, and annotation quality utilizing a multi-step decay schedule. The structure that operates is: $\{\mathrm{Ir}_{\mathrm{base}}, \{\mathrm{milestone}, \mathrm{milestone}\}, \mathrm{t_i}\}$. We consistently use a basic learning rate of 0.01 for every dataset. The quantity of training iterations and learning rate decline milestones, however, are tailored to each dataset according to its complexity and size. Specifically:

- Cityscapes perform 50 K iterations of training, with learning rate decays occurring at 32 and 44 K.
- With 144 and 176 K decays, COCO is trained for 192 K iterations.
- KITTI is trained via decay steps at 16 and 22 K over an overall of 25 K iterations.
- IDD is trained with decays of 108 and 130 K throughout 144 K cycles.

In various training setups, this staged learning rate schedule leads to better optimization and consistent convergence. A first warm-up phase is used to ensure stable convergence throughout early stages. During the first 200 iterations, the learning rate increases linearly from $\frac{1}{3}\cdot$ Ir_{base} to full Ir_{base}. The model is refined for an additional ten epochs after the initial training phase, with a fixed learning rate of $\text{lr} = 10^{-4}$. To stabilize feature distribution across GPUs, iABN sync layers are frozen during this stage of training. To quicken up computation and improve memory economy, we use mixed-precision training. Every five epochs, the model is evaluated, and early halting begins in reaction to stagnation in the Panoptic Quality (PQ) metric in order to prevent over fitting.

$$L_{total} = L_{semantic} + L_{instance} + L_{panoptic} \tag{16}$$

By making sure every single part of the segmentation pipeline is optimized at the same time, this composite loss formulation allows the model to perform well on instance, panoptic, and semantic segmentation tasks. The backbone architecture is initialized for EfficientNet-B7 with Bi-FPN, which optimizes accuracy and efficiency in computing. The Bi-FPN is included to enhance multi-scale feature fusion, helping the model to properly integrate low-resolution and high-resolution details. The model is more capable of identifying things at various sizes for the reason of its integration, making it particularly suitable for tasks like segmentation and object detection. The learning rate schedule keeps being followed, and SGD is used to optimize the ultimate loss L_{total}. The training process is the same. The efficiency of EfficientNet-B7 and the complex feature fusion capabilities of Bi-FPN are used to support the model training persistently until convergence. With the integration of EfficientNet, Bi-FPN, and panoptic segmentation, our model's domain-agnostic framework facilitates easy adaptation to non-driving conditions. Beyond autonomous driving, we are certain the model's architecture can be efficiently scaled to function well in many different kinds of applications.

4.3 Ablation Study

An ablation study for Bi-FPN with an efficient net B7 backbone that integrates instance, semantic, and panoptic segmentation evaluates the contributions of different model components using datasets like Cityscapes and COCO. As the more advanced variant of the Efficient-Net family, EfficientNet-B7 delivers greater precision and feature extraction capabilities over EfficientNet-B5, making it suited for semantic and panoptic segmentation when used together with Bi-FPN. It enhances instance segmentation by precisely determining object boundaries, semantic segmentation by correctly identifying each pixel, and panoptic segmentation by easily integrating both according to its deep architecture and high-resolution processing, which additionally boosts fine-detail recognition. EfficientNet-B7 with Bi-FPN works superior in segmentation and object detection on benchmark datasets including COCO and Cityscapes, with a higher mean average precision (mAP) in handling fragile objects and complicated urban landscapes. Using the COCO dataset, which has a variety of objects with different scales and occlusions, the suggested model improves feature fusion, which increases the accuracy of instance segmentation by efficiently distinguishing small and overlapping objects. The model incorporates exact structural information, boosting both semantic and panoptic segmentation performance, and significantly boosts segmentation consistency on the Cityscapes dataset, which focuses on urban street scenes with substantial traffic and fine-grained semantic classes. The enhanced feature extraction enables simpler recognition of cars, pedestrians, and roadways in busy regions. Faster inference times and acceptable accuracy enable EfficientNet-B5 to take a balanced approach with restricted resources for processing. EfficientNet-B7, on the other hand, offers exceptional precision and extensive feature extraction by using a deeper and more detailed architecture, which greatly enhances performance. It operates outstandingly in difficult situations and complex segmentation tasks, catching fine-grained facts that EfficientNet-B5 cannot adapt to. The depth, width, and resolution that EfficientNet-B7 improves beyond B5 enable it to capture finer-grained and more complete data, which is an essential characteristic for realizing small, restricted, and diversified objects in urban settings. Bi-FPN integration enhances the merging of multi-scale features. Consistent rises in PQ and segmentation quality are shown by empirical results across several datasets (COCO, KITTI, Cityscapes, and IDD), which support the architectural enhancements. Improved performance is offered by EfficientNet-B7 for high-accuracy applications where accuracy is vital, such as autonomous driving. Applications demanding exceptional precision in dynamic, challenging conditions are readily addressed in real-time by EfficientNet-B7 by utilizing the strength of its architecture. As demonstrated in Table 1, both the EfficientNet-B5 and B7 variants showed consistent gains in PQ, AP, and m-IOU metrics when SE layers are eliminated. The performance enhancements remain regular and low. As a result, SE layers did not feature in the final network structure.

Table 1: The ablation findings from the study for panoptic segmentation on the Cityscapes validation dataset are shown

Model	Encoder	SE	Bi-FPN	SIH	SH	PFM	PQ	PQTh	PQSt	AP	mIOU
M1	Eff-B5	✗	✓	✗	✗	✓	59.7	54.7	63.3	34.1	76.3
M2	Eff-B5	✓	✓	✗	✗	✓	61.5	57.2	64.6	36.8	77.3
M3	Eff-B5	✓	✓	✓	✓	✓	63.9	60.7	66.2	38.3	79.3
M4	Eff-B7	✗	✓	✗	✗	✓	60.4	55.1	64.2	35	77
M5	Eff-B7	✓	✓	✗	✗	✓	62.4	57.6	65.4	37.2	78.4
M6	Eff-B7	✓	✓	✓	✓	✓	64.2	61	66.5	39	80.4

Note: "St" and "Th" are superscripts that stand in the "stuff" and "thing" classes, respectively. The influence of SE removal, Bi-FPN, and fusion modules for different encoder kinds appears in the table.

4.4 Comparisons on Cityscapes Dataset

The proposed approach is demonstrated in Table 2, and it is superior through contrasting the performance of several segmentation techniques. PQ values of 56.5%, 59.0%, and 59.7% are attained by Deeper Lab [22], Adapt IS [23], and Panoptic Deep Lab [5], respectively. Solutions ranging from 55.1% to 58.8% included MTN Panoptic [12], FPS Net [14] and Prototype Panoptic [24]. Performance is enhanced by Attention PS versions, which have PQ values of 59.3% and 59.7%. Strong conditions are set by effective PS(S) and PS(M), which have PQ values of 63.9% and 65.1%, respectively. Recent models that have excellent panoptic segmentation performances include Mask-PNet [30], CCPS-Net [31], and Efficient PS-B4-RCC [32]. PQ is 61.8% for Mask-PNet with ResNet-50, 60.5% for CCPS-Net (ResNet-50), and 61.2% for ResNet-101, and PQ is 64.2% for Efficient PS-B4-RCC. PQ scores of 64.2% and 65.5% are attained by the suggested methods "Proposed (S)" and "Proposed (M)", outperforming the current methods with commensurate gains in PQTh and PQSt. These outcomes confirm the efficacy of the indicated strategies, revealing their versatility and potential for extremely successful panoptic segmentation. Despite the modest PQ enhancement on Cityscapes (65.5% vs. 65.1%), benchmark datasets often contain higher-resolution, cleaner images with fewer fluctuations than real-time scenarios. More robustness over dynamic, real-life conditions can be the result of even little improvements in such controlled datasets. Consequently, the observed advantages suggest a significant improvement in performance under operational challenges.

Table 2: Panoptic segmentation performance comparison on the Cityscapes validation dataset as described

Methods	Backbone	PQ (%)	PQ^{Th} (%)	PQ^{St} (%)
Deeper lab	Xception-71	56.5	–	–
AdaptIS	ResNet 50	59.0	55.8	61.3
Panoptic deep lab	ResNet 50	59.7	–	–
MTN panoptic	ResNet 50-FPN	57.3	53.9	59.7
FPS net	ResNet 50-FPN	55.1	48.3	60.1
Prototype panoptic	VolvNet2-39-FPNLite	57.3	50.4	62.4
Attention PS	ResNet 50-FPN Lite	59.3	52.8	64.1
Attention PS	VolvNet2-39-FPNLite	59.7	52.8	64.7
Efficient PS(S)	Efficient-Net B5	63.9	60.7	66.2
Efficient PS(M)	Efficient-Net B5	65.1	61.5	67.7
Mask-P net	ResNet 50	61.8	59.4	64.0
CCPS net	ResNet 50	60.5	56.9	63.1
CCPS net	ResNet 101	61.2	57.1	64.1
Efficient PS-b4-RCC	Efficient-Net B4	64.2	59.8	67.6
Proposed (S)	Efficient-Net B7	64.2	61.0	66.5
Proposed (M)	Efficient-Net B7	65.5	61.9	67.9

Note: Superscripts St and Th refer to 'stuff' and 'thing' classes; S and M refer to one scale and multi-scale, respectively.

4.5 Comparisons on the COCO Dataset

The success of the proposed approach is shown by the performance comparison of the segmentation methods in Table 3. The relative PQ scores of Deeper Lab [22], Adapt IS [23], and Panoptic Deep Lab [5] are 33.8%, 34.4%, and 35.1%. Scores increase to 37.5% and 37.1% for PCV [27], respectively. DetR [12] and Panoptic FCN [29] all work together to improve PQ, which ranges from 43.0% to 43.8%. The PQ scores of complex techniques such as Center-Guided PS [21], Mask2Former [9], and Panoptic Seg Former [6] range from 49.6% to 52.2%. The Panoptic Quality (PQ) score of CCPS-Net [31] is 43.0% while employing a ResNet-50 backbone and 43.5% while implementing a ResNet-101 backbone, signifying major improvements in PQ. The lower PQ score of 37.3% is all that Mask-PNet [30] with ResNet-50 tries to achieve. Superior segmentation performance and resilience are shown by the proposed approach, which surpasses all others with a PQ of 52.4%, PQ^{Th} of 58.9%, and PQ^{St} of 42.8%.

Table 3: Panoptic segmentation performance comparison on the COCO validation dataset is described below

Methods	Backbone	PQ (%)	PQ^{Th} (%)	PQ^{St} (%)
Deeper lab	Xception-71	33.8	–	–
Adapt IS	ResNet 50	34.4	50.0	29.3
Panoptic deep lab	ResNet 50	35.1	–	–

(Continued)

Table 3 (continued)

Methods	Backbone	PQ (%)	PQ^{Th} (%)	PQ^{St} (%)
PCV	ResNet 50	37.5	40.7	33.1
Unifying	ResNet 50-FPN	43.4	48.6	35.5
Single-shot-576	ResNet 50-FPN	32.4	34.8	28.6
Attention PS-800	ResNet 50-FPN Lite	34.4	39.3	27.1
Attention PS-640	ResNet 50-FPN Lite	33.4	37.8	26.7
DetR	ResNet 50	43.2	48.2	36.1
Panoptic FCN	ResNet 50	43.6	49.3	35.0
Panoptic seg former	ResNet 50	49.6	54.4	42.4
Mask2former	ResNet 50	51.9	57.7	43.0
Center-Guided PS	ResNet 50	52.2	58.4	42.6
Mask-PNet	Resnet 50	37.3	–	–
CCPS net	ResNet 50	43.0	49.2	33.6
CCPS net	ResNet 101	43.5	49.9	33.8
Proposed	Efficient-Net B7	52.4	58.9	42.8

Note: Superscripts St and Th refer to 'stuff' and 'thing' classes, respectively.

4.6 Comparisons on KITTI Dataset

The efficiency of the recommended strategy is shown in Table 4 by the segmentation methods' performance evaluation. In the single-scale mode, Panoptic FPN [3] and Seamless [11] attain PQ scores that vary from 38.6% to 41.3%, while Efficient PS attains 42.9%. The recommended method outperformed all others, achieving 73.1% SQ, 54.1% RQ, and 43.3% PQ. In the multi-scale mode, Efficient PS attains 43.7% PQ and Efficient PS-b4 RCC 43.3%, while Panoptic FPN, UPS Net, and Seamless vary from 39.3% to 42.2% PQ. With 44.0% PQ, 73.7% SQ, and 54.5% RQ, the suggested method performs better than the others, indicating its increased segmentation accuracy. It also attains the highest mIoU (56.7) and AP (28.3), showing its adaptability and efficacy in high-performance segmentation tasks.

Table 4: Panoptic segmentation performance comparison on the KITTI validation dataset is shown below

Methods	PQ (%)	SQ (%)	RQ (%)	PQ^{Th} (%)	SQ^{Th} (%)	RQ^{Th} (%)	PQ^{Th} (%)	SQ^{Th} (%)	RQ^{Th} (%)	AP	mIOU
					Single-scale (Mode)						
Panoptic FPN	38.6	70.4	51.2	26.1	68.3	40.1	47.6	71.9	59.2	24.4	52.1
Seamless	41.3	71.7	52.3	28.5	69.2	42.3	50.6	73.6	59.6	24.9	53.8
Efficient PS	42.9	72.7	53.6	30.4	69.8	43.7	52.0	74.9	60.9	27.1	55.3
Proposed	43.3	73.1	54.1	30.7	70.1	44.1	52.4	75.3	61.3	27.5	55.6
					Multi-scale (Mode)						
Panoptic FPN	39.3	70.8	51.6	26.9	68.7	40.4	48.3	72.4	59.8	24.8	52.8
Seamless	42.2	72.3	52.9	29.1	69.7	42.9	51.8	74.2	60.1	26.6	55.1

(Continued)

Table 4 (continued)

Methods	PQ (%)	SQ (%)	RQ (%)	PQ^{Th} (%)	SQ^{Th} (%)	RQ^{Th} (%)	PQ^{Th} (%)	SQ^{Th} (%)	RQ^{Th} (%)	AP	mIOU
Efficient PS	43.7	73.2	54.1	30.9	70.2	44.0	53.1	75.4	61.5	27.9	56.4
Efficient PS-b4-RCC	43.3	74.8	53.6	30.1	69.8	43.1	52.6	75.1	61.1	–	–
Proposed	44.0	73.7	54.5	31.6	70.6	44.4	53.7	75.7	61.7	28.3	56.7

Note: Superscripts St and Th refer to 'stuff' and 'thing' classes, respectively.

4.7 Comparisons on IDD Dataset

The proposed method performed better in both one-scale and various-scale modes on every evaluation criterion shown in Table 5. The highest Panoptic Quality (PQ) of 50.4%, Segmentation Quality (SQ) of 78.8%, and Recognition Quality (RQ) of 62.4% are attained in the single-scale setting. In the "thing" (PQ^{Th}: 51.0%) and "things" (PQ^{St}: 50.1%) categories, it likewise performs higher than other methods. In the multi-scale mode, the proposed strategy produces the maximum PQ of 51.6%, SQ of 79.1%, and RQ of 63.9%. Furthermore, it shows the best results in the segmentation of things (PQ^{Th}: 52.7%) and stuff (PQ^{St}: 50.4%). Furthermore, it demonstrates its efficacy in precision and excellent panoptic segmentation by attaining the highest mean intersection over union (mIoU: 72.5) and average precision (AP: 33.0).

Table 5: Panoptic segmentation performance comparison on the IDD validation dataset

Methods	PQ (%)	SQ (%)	RQ (%)	PQ^{Th} (%)	SQ^{Th} (%)	RQ^{Th} (%)	PQ^{St} (%)	SQ^{St} (%)	RQ^{St} (%)	AP	mIOU
Single-scale (Mode)											
Panoptic FPN	45.9	75.9	60.8	46.1	77.8	60.9	45.8	74.9	60.7	27.8	68.1
Seamless	47.7	77.2	61.2	48.9	79.5	61.5	47.1	76.1	61.1	30.1	69.6
Efficient PS	50.1	78.4	62.0	50.7	80.6	61.6	49.8	77.1	62.2	31.6	71.3
Proposed	50.4	78.8	62.4	51.0	80.8	61.8	50.1	77.4	62.5	31.9	71.6
Multi-scale (Mode)											
Panoptic FPN	46.7	77.0	61.0	47.3	78.9	61.1	46.4	76.1	61.0	28.9	70.1
Seamless	48.5	78.2	61.9	49.5	80.4	62.2	47.9	77.1	61.7	31.4	71.3
Efficient PS	51.1	78.8	63.5	52.6	81.2	65.4	50.3	77.5	62.5	32.9	72.1
Efficient PS-b4-RCC	51.2	78.9	64.4	50.2	80.3	61.6	52.1	79.8	64.8	–	–
Proposed	51.6	79.1	63.9	52.7	81.5	65.6	50.4	77.9	62.7	33.0	72.5

Note: 'Stuff' and 'thing' classes are denoted by superscripts St and Th, respectively.

With an input size of 1025 × 2048 pixels, the suggested method has 435.37 B FLOPs, 42.21 M parameters, and a processing time of 170 ms. While keeping processing time similar, the suggested solution slightly increases computational complexity in comparison to Efficient PS, which has 40.89 M parameters and 433.94 B FLOPs. With 51.43 M parameters and 514.00 B FLOPs, Seamless has a decreased inference time of 168 ms but a greater calculating load. UPS Net has the quickest processing time, at 202 ms, with 45.05 M parameters and 487.02 B FLOPs equally. This comparison illustrates how successfully the proposed framework strikes an equilibrium between performance and computational cost. The suggested strategy delivers a remarkable level of segmentation accuracy, which is crucial for credible perception in autonomous driving, despite a little speed lag. In difficult circumstances, this trade-off favors safer and more accurate detection, and additional speed optimization is possible with hardware acceleration and model compression techniques. Although the use of EfficientNet-B7 and Bi-FPN improves the computational demand of the model, this framework promises high precision and thorough segmentation—two essential elements for safety-critical jobs in autonomous driving. When targeting platforms with few assets, the architecture's adaptability provides future modification with the use of lightweight backbones or pruning techniques, ensuring scalability without compromising speed.

4.8 Comparison of Panoptic Quality on Four Dataset

Panoptic Quality (PQ) is used to assess the performance measures across four datasets: Cityscape, COCO, KITTI, and IDD as shown in Fig. 5. The Panoptic Quality (PQ) of the proposed method is 65.5% on Cityscapes, 52.4% on COCO, 44% on KITTI, and 51.6% on IDD, demonstrating its outstanding efficacy across a number of datasets. The model's ability to understand urban scenes is demonstrated by its greatest PQ score on Cityscapes; nevertheless, problems with challenging autonomous driving scenarios are suggested by its poorer performance on KITTI. Its capacity to be extrapolated across numerous real-world datasets is confirmed by the COCO and IDD results.

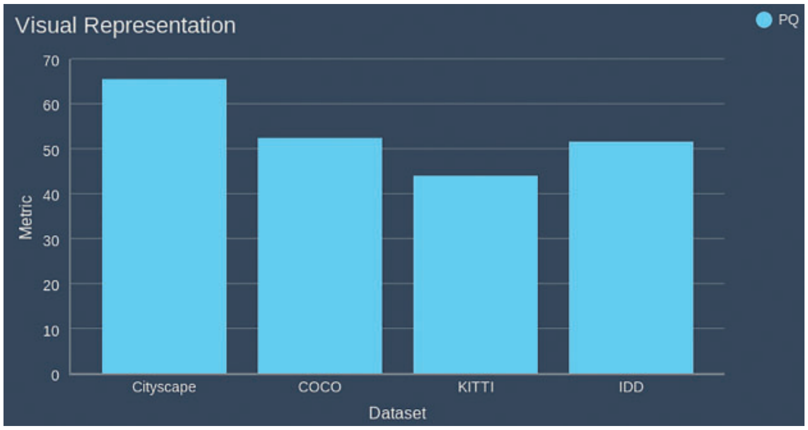


Figure 5: Comparison of panoptic quality on Cityscapes, COCO, KITTI, and IDD validation dataset

4.9 Visualization

The effectiveness of the proposed approach on a variety of datasets is demonstrated by the panoptic segmentation results. Fig. 6 shows how well it adjusts to vehicle-centric scenarios and urban surroundings using the Cityscapes and KITTI datasets. Fig. 7 further demonstrates the robustness of the technique in the dynamic and complex contexts of the COCO and IDD datasets. These results demonstrate how accurately the approach can discriminate between items and backgrounds in challenging scenarios. The method's consistent performance across several datasets attests to its reliability and generalizability for panoptic segmentation tasks.

Fig. 8 shows the qualitative analysis of the proposed model capturing distant objects on a cityscape, KITTI, COCO, and IDD validation dataset. In the majority of cases, the suggested technique surpasses the baseline in realizing small objects, showing strong capability. In scenes that are jam-packed, it can detect fine-grained details. There is still some opportunity for advancement, though, especially in cases of distant objects or strong occlusions.

In the Cityscapes, KITTI, COCO, and IDD validation datasets, challenging scenarios are highlighted in Fig. 9. These include items that overlap in scenes with a lot of traffic, everything that in far away or small-scale on broad highways, and frames taken in low light, like evening light or shadows. These instances reflect the complexity of real-life circumstances, which often result in missed detections and decreased segmentation accuracy because of occlusion, low visibility, or size fluctuation. Improving the model's capacity to collect fine-grained features under tough conditions is one area of future development that is going to keep addressing these limitations.

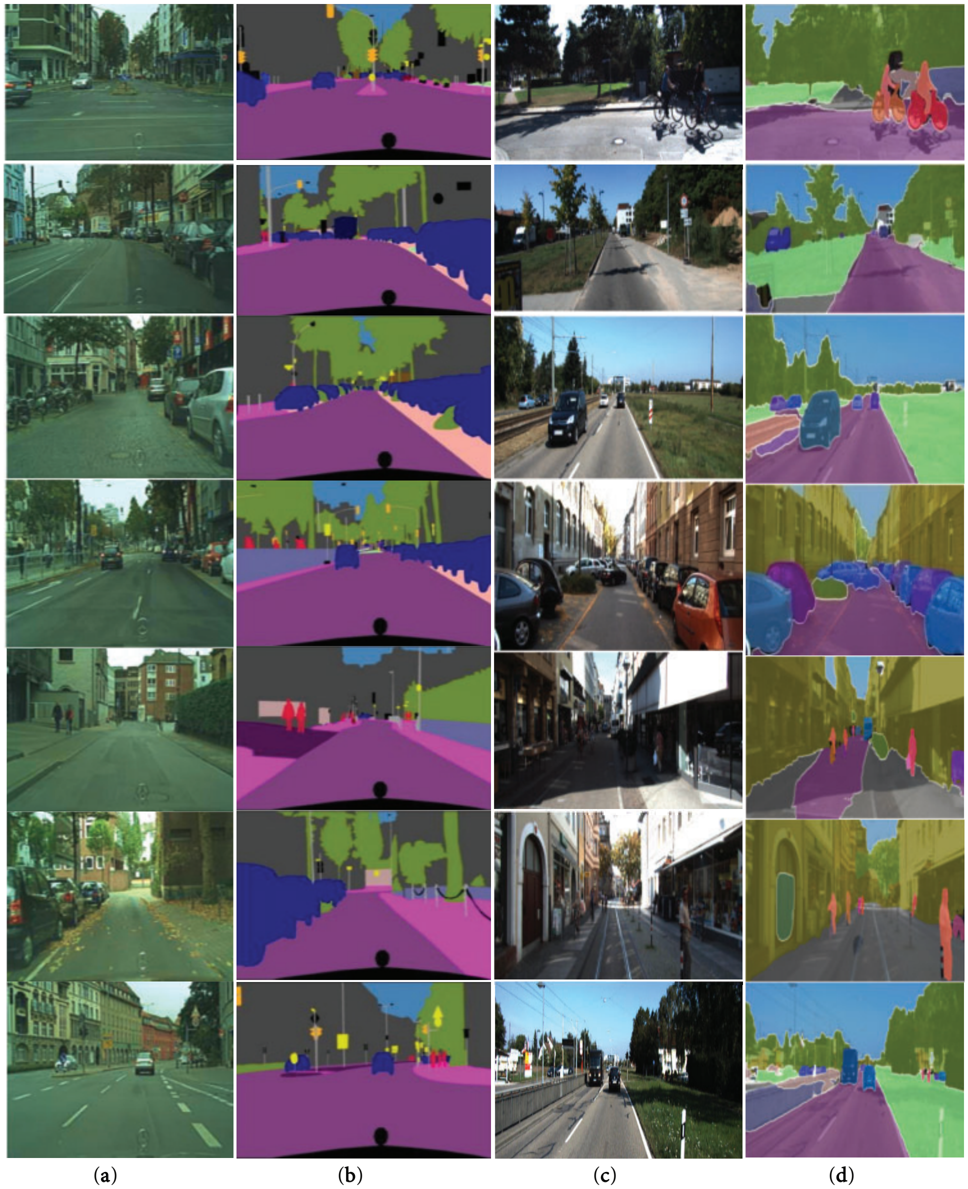


Figure 6: Panoptic segmentation performance of the recommended network can be seen using two validation datasets: (**a**) input image for Cityscapes; (**b**) suggested output for Cityscapes; (**c**) input image for KITTI; and (**d**) proposed output for KITTI

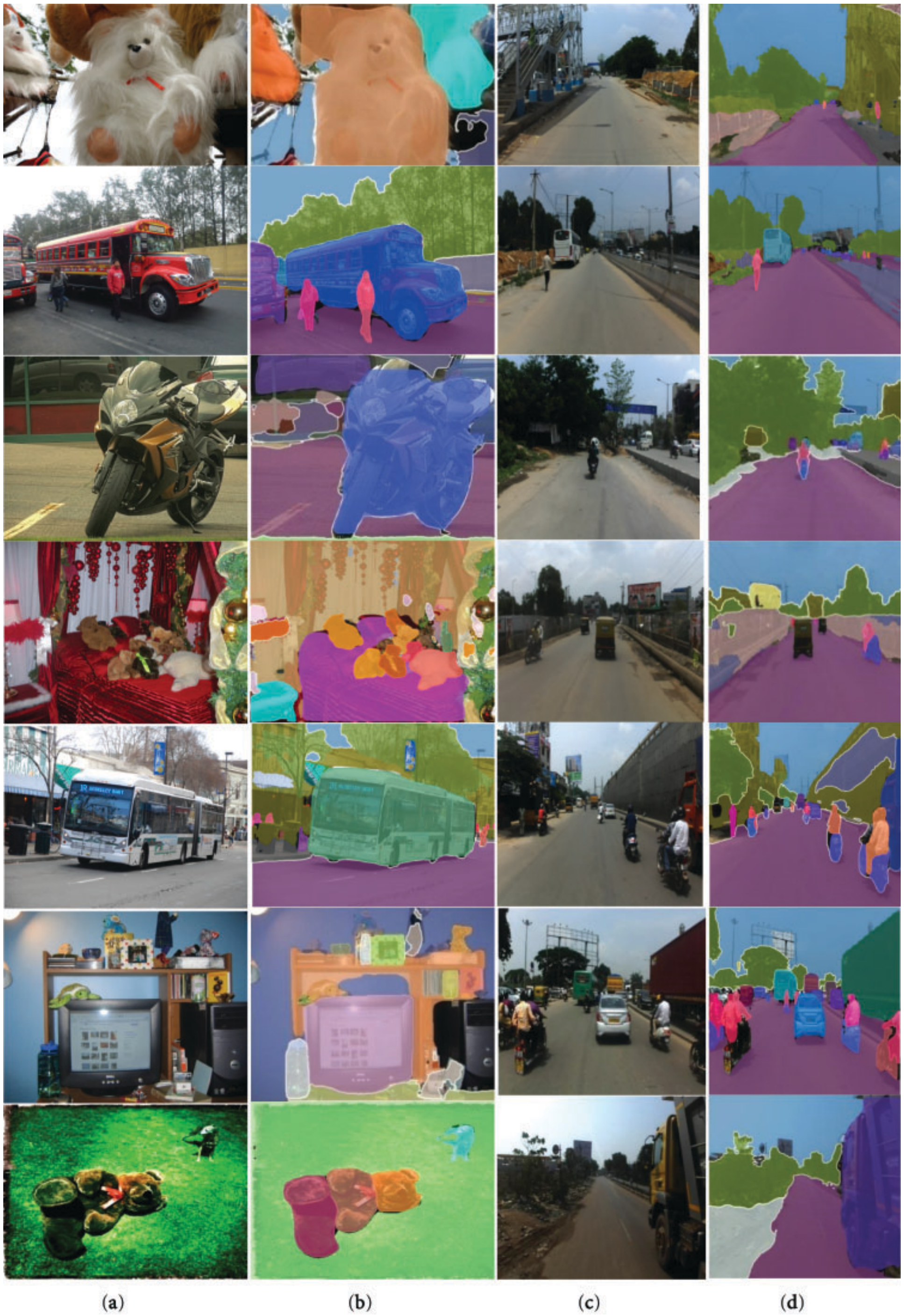


Figure 7: Visualization of the proposed network's panoptic segmentation performance using two validation datasets: (**a**) input image for COCO; (**b**) output that is suggested for COCO; (**c**) input image for IDD; and (**d**) output that is suggested for IDD

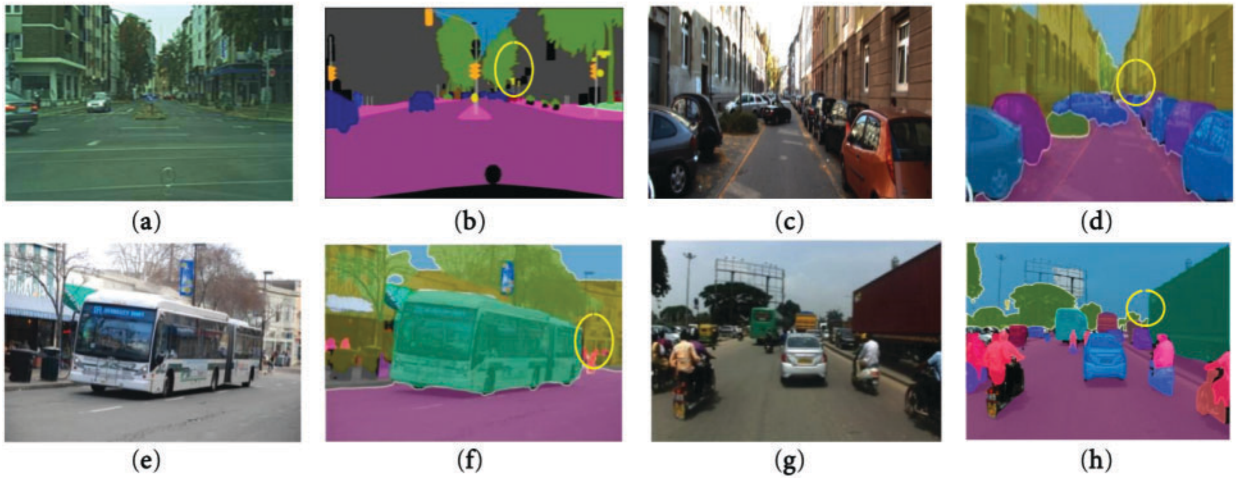
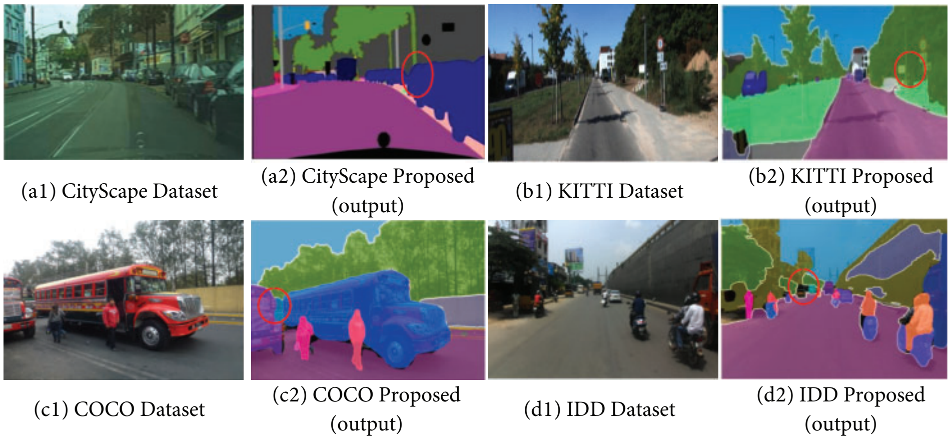


Figure 8: Evaluation of the specified network's panoptic segmentation performance qualitatively with four validation datasets: (**a**) input images for Cityscapes; (**b**) output proposed on Cityscapes; (**c**) input image of KITTI; (**d**) output proposed on KITTI; (**e**) input image of COCO; (**f**) output proposed on COCO; (**g**) input image of IDD; (**h**) output proposed on IDD



(a1) CityScape Dataset

(a2) CityScape Proposed (output)

(b1) KITTI Dataset

(b2) KITTI Proposed (output)

(c1) COCO Dataset

(c2) COCO Proposed (output)

(d1) IDD Dataset

(d2) IDD Proposed (output)

Figure 9: Challenging cases of panoptic segmentation performance results of the Proposed Network on the (**a**) City Scape and (**b**) KITTI validation dataset (**c**) COCO and (**d**) IDD validation dataset

A comparative comparison between Efficient PS and the suggested model under the same difficult circumstances is provided in Fig. 10. The ability of Efficient PS to detect smaller or less noticeable instances and segment overlapping objects is restricted. Especially in low-light or sheltered environments. In several frames, the proposed approach demonstrates slight improvements, with more thorough segmentation and substantially sharper bounds. Nevertheless, the differences are not equal in every instance, and both models exhibit positive and negative characteristics based on specific circumstances.

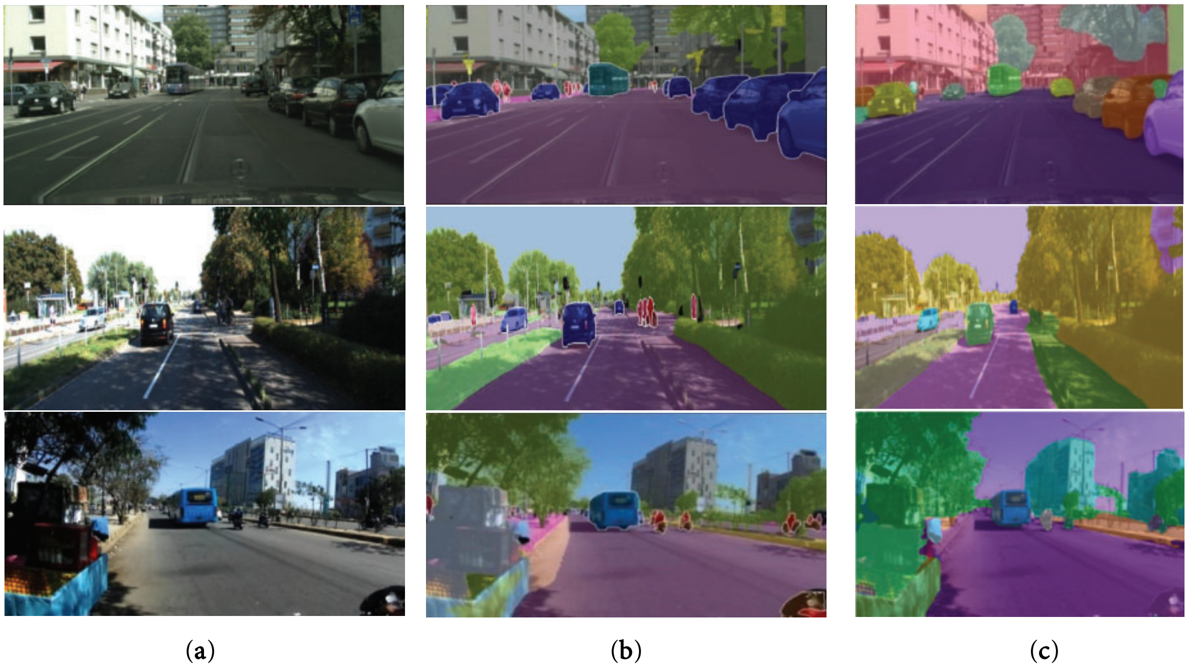


Figure 10: Comparison among the baseline model and the following datasets: (**a**) the input image from Cityscapes; (**b**) the output from Efficient PS; and (**c**) the output from the suggested approach on the Cityscapes, KITTI, and IDD validation datasets

5 Conclusion and Future Work

Using the combination of the existing panoptic segmentation and the substitution of BI-FPN for efficient net B5 and Bi-FPN for efficient net B7. We present an enhanced panoptic segmentation technique. The modified architecture outperformed the baseline in terms of segmentation accuracy and resilience, according to extensive testing on the Cityscapes, KITTI, COCO, and IDD datasets. The proposed model accurately segmented backgrounds and objects under challenging conditions, demonstrating its flexibility in a range of scenarios. However, little object detection needs to be improved because it is still quite challenging. Despite this limitation, the results show that efficient Net B7 combined with Bi-FPN is a workable framework for achieving reliable panoptic segmentation in a range of applications. Bi-FPN partially tackles the issue of small item recognition by retaining fine-grained spatial data, thereby rendering multi-scale feature fusion easier. EfficientNet-B7 additionally enables high-resolution feature extraction to assist with small object recognition. The following studies will employ datasets from different places and domain adaptation strategies to mitigate geographic bias and class imbalance. In addition, techniques to address imbalance within rare categories will be researched, such as synthetic oversampling and class-balanced loss. In the future, the model will be improved for embedded systems utilizing model compression approaches such as knowledge distillation, quantization, and pruning. For enhancing the model's responsiveness in fluctuating environments in real life, adaptive inference strategies will also be investigated. The subsequent studies will offer an extensive strategy for the implementation of adaptive tiling, which includes dynamic region choosing based on scene context, tile size optimization, and real-time benchmarking confirming efficiency improvements in a number of driving settings. Future studies will focus on improving the model's performance in demanding scenarios and the challenges of identifying small objects. Investigating efficient feature extraction techniques and thin backbone architectures that reduce computing costs without compromising accuracy can help achieve this. To enhance the model's ability to focus on crucial details, particularly tiny and hidden objects, transformers and other advanced attention mechanisms will be included. Additionally, the model will be evaluated on a greater range of real-world datasets to ensure scalability and adaptability. Priority will also be given to deployment in real-time applications to verify its utility. The system should become more effective and successful for panoptic segmentation tasks with further improvement.

Acknowledgement: Sincere thanks are expressed by the authors to Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology for providing the facilities and assistance necessary for carrying out this study.

Funding Statement: The authors received no specific funding for this study.

Author Contributions: Conceptualization, Darthy Rabecka V and Britto Pari J; methodology, Darthy Rabecka V and Britto Pari J; software, Darthy Rabecka V and Britto Pari J; validation, Darthy Rabecka V and Britto Pari J; formal analysis, Darthy Rabecka V and Britto Pari J; investigation, Darthy Rabecka V and Britto Pari J; resources, Darthy Rabecka V; data curation, Darthy Rabecka V; writing—original draft preparation, Darthy Rabecka V and Britto Pari J; writing—review and editing, Darthy Rabecka V, Britto Pari J and Man-Fai Leung; visualization, Darthy Rabecka V; supervision, Britto Pari J and Man-Fai Leung; project administration, Britto Pari J and Darthy Rabecka V. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The study's findings are available to the public at https://doi.org/10.1109/CVPR.2016.350 (accessed on 01 July 2025) in Cityscapes [20]. Public access to the study's data is available at http://panoptic.cs.uni-freiburg.de (accessed on 02 July 2025) in KITTI and IDD [16]. The public can access the study's data in the Coco dataset at https://cocodataset.org/#home (accessed on 30 June 2025).

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

1. Yan W, Qian Y, Zhuang H, Wang C, Yang M. Sam4udass: when sam meets unsupervised domain adaptive semantic segmentation in intelligent vehicles. IEEE Trans Intell Veh. 2023;9(2):3396–408. doi:10.1109/tiv.2023.3344754.
2. Shokri D, Larouche C, Homayouni S. Proposing an efficient deep learning algorithm based on Segment Anything Model for detection and tracking of vehicles through uncalibrated urban traffic surveillance cameras. Electronics. 2024;13(14):2883. doi:10.3390/electronics13142883.
3. Li Y, Zhao H, Qi X, Wang L, Li Z, Sun J, et al. Fully convolutional networks for panoptic segmentation. In: Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2021 Jun 20–25; Nashville, TN, USA.
4. Chamola V, Chougule A, Sam A, Hussain A, Yu FR. Overtaking mechanisms based on augmented intelligence for autonomous driving: data sets, methods, and challenges. IEEE Internet Things J. 2024;11(10):17911–33. doi:10.1109/jiot.2024.3362851.
5. Cheng B, Collins MD, Zhu Y, Liu T, Huang TS, Adam H, et al. Panoptic-deeplab: a simple strong and fast baseline for bottom-up panoptic segmentation. In: Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2020 Aug 5; Seattle, WA, USA.
6. Audinys R, Šlikas Ž, Radkevičius J, Šutas M, Ostreika A. Deep reinforcement learning for a self-driving vehicle operating solely on visual information. Electronics. 2025;14(5):825. doi:10.3390/electronics14050825.
7. Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A, Zagoruyko S. End-to-end object detection with transformers. In: Proceedings of the Computer Vision—ECCV 2020: 16th European Conference; 2020 Aug 23–28; Glasgow, UK.
8. Cheng B, Schwing A, Kirillov A. Per-pixel classification is not all you need for semantic segmentation. Adv Neural Inf Process Syst. 2021;34:17864–75.

9. Cheng B, Misra I, Schwing AG, Kirillov A, Girdhar R. Masked-attention mask transformer for universal image segmentation. In: Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2022 June 18–24; Orleans, LA, USA.
10. Feng D, Haase-Schütz C, Rosenbaum L, Hertlein H, Glaeser C, Timm F, et al. Deep multi-modal object detection and semantic segmentation for autonomous driving: datasets, methods, and challenges. IEEE Trans Intell Transp Syst. 2020;22(3):1341–60. doi:10.1109/tits.2020.2972974.
11. Porzi L, Rota Bulo S, Colovic A, Kontschieder P. Seamless scene segmentation. In: Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2019 June 15–20; Long Beach, CA, USA.
12. Petrovai A, Nedevschi S. Multi-task network for panoptic segmentation in automated driving. In: Proceedings of the 2019 IEEE Intelligent Transportation Systems Conference (ITSC); 2019 Oct 27–30; Auckland, New Zealand.
13. Hu S, Fang Z, Deng Y, Chen X, Fang Y, Kwong S. Toward full-scene domain generalization in multi-agent collaborative bird's eye view segmentation for connected and autonomous driving. IEEE Trans Intell Transp Syst. 2024;26(2):1783–96. doi:10.1109/tits.2024.3506284.
14. De Geus D, Meletis P, Dubbelman G. Fast panoptic segmentation network. IEEE Robot Autom Lett. 2020;5(2):1742–9. doi:10.1109/lra.2020.2969919.
15. Ahmed R, Elsayed Mahmoud Salaheldin, Abd-Elkawy Eman H. A novel hybrid deep learning algorithm for object and lane detection in autonomous driving. JAIT [Internet]. 2025 May 18 [cited 2025 Jul 1]. Available from: https://ojs.istp-press.com/jait/article/view/695.
16. Mohan R, Valada A. EfficientPS: efficient panoptic segmentation. Int J Comput Vis. 2021;129(5):1551–79. doi:10.1007/s11263-021-01445-z.
17. Zhang S, Ni P, Wen J, Han Q, Du X, Xu K. Non-contact impact load identification based on intelligent visual sensing technology. Struct Health Monit. 2024;23(6):3525–44. doi:10.1177/14759217241227365.
18. Zhang S, Ni P, Wen J, Han Q, Du X, Xu K. Automated vision-based multi-plane bridge displacement monitoring. Autom Constr. 2024;166:105619. doi:10.1016/j.autcon.2024.105619.
19. Zhang S, Ni P, Wen J, Han Q, Du X, Fu J. Intelligent identification of moving forces based on visual perception. Mech Syst Signal Process. 2024;214:111372. doi:10.1016/j.ymssp.2024.111372.
20. Petrovai A, Nedevschi S. Fast panoptic segmentation with soft attention embeddings. Sensors. 2022;22(3):783. doi:10.3390/s22030783.
21. Baek JH, Lee HK, Choo HG, Jung SH, Koh YJ. Center-guided transformer for panoptic segmentation. Electronics. 2023;12(23):4801. doi:10.3390/electronics12234801.
22. Yang TJ, Collins MD, Zhu Y, Hwang JJ, Liu T, Zhang X, et al. Deeperlab: single-shot image parser. arXiv:1902.05093. 2019.
23. Sofiiuk K, Barinova O, Konushin A. Adaptis: adaptive instance selection network. In: Proceedings of 2019 IEEE/CVF International Conference on Computer Vision (ICCV); 2019 Oct 27–Nov 2; Seoul, Republic of Korea.
24. Petrovai A, Nedevschi S. Real-time panoptic segmentation with prototype masks for automated driving. In: Proceedings of the 2020 IEEE Intelligent Vehicles Symposium (IV); 2020 Oct 19–Nov 13; Las Vegas, NV, USA.
25. Yang M. Application of deep learning in autonomous driving. Appl Comput Eng. 2025;145:135–40.
26. Zhang Y. Research on training optimization of artificial intelligence autonomous driving model based on big data analysis. Appl Comput Eng. 2025;141:254–60. doi:10.54254/2755-2721/2025.22041.
27. Weber M, Luiten J, Leibe B. Single-shot panoptic segmentation. In: Proceedings of the 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS); 2020 Oct 24–Jan 24; Las Vegas, NV, USA.
28. Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, et al. Microsoft COCO: common objects in context. In: Proceedings of the Computer Vision—ECCV 2014: 13th European Conference; 2014 Sep 6–12; Zurich, Switzerland.
29. Lei Z, Wang W, Zhu Z, Ma J, Ge SS. Safe motion planning for multi-vehicle autonomous driving in uncertain environment. IEEE Robot Autom Lett. 2025;10(3):2199–206. doi:10.1109/lra.2025.3528254.
30. Xian PF, Po LM, Xiong JJ, Zhao YZ, Yu WY, Cheung KW. Mask-pyramid network: a novel panoptic segmentation method. Sensors. 2024;24(5):1411. doi:10.3390/s24051411.

31. Xu Y, Liu R, Zhu D, Chen L, Zhang X, Li J. Cascade contour-enhanced panoptic segmentation for robotic vision perception. Front Neurorobot. 2024;18:1489021. doi:10.3389/fnbot.2024.1489021.
32. Benkirane FE, Crombez N, Hilaire V, Ruichek Y. Hybrid AI for panoptic segmentation: an informed deep learning approach with integration of prior spatial relationships knowledge. Comput Vis Image Underst. 2024;240:103909. doi:10.1016/j.cviu.2023.103909.