Journal of Information Systems Engineering and Management
2025, 10(4s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Semantic Object Segmentation using Modified HRNet Deep Learning Model

Abdul Saleem L[1], Dr. Gowtham Mamidisetti[2]

[1]Research Scholar, Dept. of Computer Science & Engineering, Malla Reddy University, Hyderabad, India.

skabdulsaleem@gmail.com

[2]Professor, Dept. of Computer Science & Engineering, Malla Reddy University, Hyderabad, India.

drmgowtham@mallareddyuniversity.ac.in

ARTICLE INFO

Received: 12 Oct 2024

Revised: 11 Dec 2024

Accepted: 24 Dec 2024

ABSTRACT

This paper presents a modified High-Resolution Network (HRNet) architecture designed for enhanced semantic object segmentation, addressing the prevalent challenges of accuracy and computational efficiency in complex image analyses. Through strategic modifications to the conventional HRNet, including optimized feature blocks and advanced feature extraction techniques, our model demonstrates significant improvements in segmentation performance. Utilizing the Cityscapes dataset, renowned for its comprehensive urban scene representation, our enhanced HRNet model achieved a notable increase in validation accuracy to 85.8% and mean Intersection over Union (mIoU) to 63.43%, surpassing the original HRNet benchmarks by 3.39% and 3.43%, respectively. These results highlight not only the precision of our model in delineating intricate urban landscapes but also its robustness in handling diverse object scales and complexities. The qualitative analysis, underscored by comparison images, reveals our model's ability to produce more defined segmentation contours and accurate object identifications, setting a new standard for HRNet-based semantic segmentation. This work not only advances the field of high-resolution image segmentation but also offers a foundational model for future research aimed at solving increasingly complex image processing challenges

Keywords: semantic segmentation, High-Resolution Network (HRNet), feature pyramids, high-resolution feature maps

INTRODUCTION

Every pixel in an image must be identified and given a class name as part of semantic segmentation. By identifying various objects and their borders, it seeks to offer a thorough grasp of an image's visual content. A modified High-Resolution Network (HRNet) deep learning model is presented in this study for semantic picture segmentation [1]. The goal of object segmentation is to identify the boundaries of objects and assign a unique label to each segmented region, enabling machines to understand the visual content of an image in a more sophisticated way. Object segmentation has numerous applications, including autonomous navigation, robotics, medical imaging, and augmented reality[2]. Deep learning-based approaches have recently revolutionized the area of object segmentation by enabling very precise and effective segmentations even in crowded and complicated environments [3]. This technology is rapidly advancing and holds great potential for enhancing the capabilities of computer vision systems in various domains.

In recent years, owing to the development of approaches that are based on deep learning, object segmentation methods have seen considerable improvements in both their accuracy and their speed. These advancements have made object segmentation more reliable and accessible, opening up new possibilities for innovation and research in computer vision [4]. Hence, object segmentation has become a crucial aspect of modern computer vision systems, and its importance is expected to continue to grow in the future. Deep learning models have been widely used for object segmentation in recent years [5], which has increased the popularity of this technology. Deep learning models have the ability to achieve state-of-the-art performance in a number of computer vision applications. Deep learning models are capable of learning complex and hierarchical representations of visual features from large-scale datasets, enabling them to recognize and segment objects accurately and efficiently [6].

In spite of the substantial advancements that have been achieved in object segmentation via the use of deep learning models [7], there are still a number of research gaps that must be addressed in order to increase the accuracy, efficiency, as well as generalisation of these models. Some of the main research gaps in object segmentation using deep learning models are: Deep learning models for object segmentation are susceptible to noise and occlusion, which can significantly affect the accuracy of the segmentation [8]. There is a need for new techniques that can handle noise and occlusion effectively and produce accurate segmentation results. Deep learning models for object segmentation are often trained on large-scale datasets that may not represent the target domain accurately. There is a need for new techniques that can improve the generalization of these models to unseen data and domain shifts. Because deep learning models for object segmentation are sometimes seen as "black boxes," it might be difficult to comprehend how these models get to the judgements that they do about segmentation. There is a need for innovative methods that may give explanations as well as insights into the process of segmentation, hence boosting the interpretability as well as trustworthiness of these models. These models can be improved by developing new methods.

In the realm of semantic segmentation, the High-Resolution Network (HRNet) has emerged as a pivotal architecture, known for its capability to maintain high-resolution representations through the network, thus facilitating precise object delineation. However, despite its successes, HRNet and similar models face inherent limitations, particularly when processing complex urban scenes with varying object scales and densely populated areas. These challenges include a notable sensitivity to noise and occlusion, a tendency towards computational intensity, and a struggle with the diverse and dynamic nature of urban environments, leading to suboptimal segmentation accuracy and efficiency. Addressing these critical gaps, our study introduces a refined HRNet architecture, meticulously engineered to bolster segmentation precision while curbing computational demands. By optimizing feature extraction blocks and integrating advanced feature fusion techniques, we significantly enhance the model's robustness against the multifaceted challenges presented by urban scenes. Our modifications aim to transcend the traditional boundaries of HRNet's performance, offering a more versatile and efficient solution for semantic segmentation tasks across varied and intricate environments

Deep learning has revolutionized the field of object segmentation [9], offering unparalleled accuracy and the ability to process complex images with varying object scales and complexities. However, the implementation of deep learning algorithms for object segmentation presents substantial challenges, primarily due to their considerable demand for extensive training data and significant computational resources. This demand results in computationally intensive and time-consuming training processes, exacerbating the challenge in fields where annotated data are scarce, such as medical imaging. Consequently, there is an urgent need for innovative techniques that enhance training efficiency and data annotation processes without compromising model performance.

In response to these challenges, this paper proposes a modified High-Resolution Network (HRNet) architecture that introduces several strategic improvements aimed at advancing the current state-of-the-art in semantic segmentation. The objectives of this work are outlined as follows:

- **Optimization of Feature Extraction:** We introduce optimized feature blocks to enhance the model's ability to integrate detailed image characteristics seamlessly. This modification aims to improve segmentation precision across diverse object scales and complexities.
- **Advanced Feature Fusion Techniques:** By implementing sophisticated feature fusion techniques, our model effectively handles multi-resolution information. This enhancement ensures refined segmentation accuracy for both large and small objects within complex urban scenes.
- **Computational Efficiency:** Addressing the computational intensity traditionally associated with HRNet, we streamline the architecture. This approach significantly reduces computational demands without sacrificing segmentation performance.
- **Robustness Against Environmental Variabilities:** Our model is enhanced to withstand environmental factors such as noise and occlusion, prevalent in dynamic urban environments, thus ensuring consistent segmentation accuracy under varied conditions.
- **Versatility and Scalability:** The modified HRNet architecture is designed to be versatile and scalable, capable of adapting to a wide range of segmentation tasks and datasets beyond urban scenes, including medical imaging and aerial photography.

The necessity for such advancements is underscored by the continuous evolution of deep learning-based object segmentation, which has emerged as a key area of research due to its potential to unlock new and exciting applications in computer vision. This potential is fueled by the availability of large-scale datasets and the development of innovative architectures and optimization techniques. Our work introduces an effective deep learning model tailored to precisely segment objects. By leveraging HRNet's ability to handle high-resolution inputs efficiently and maintaining computational efficiency, our model processes parallel streams at various resolutions, capturing both fine-grained features and broad contextual information. Unlike traditional approaches that compromise resolution for computational efficiency, our HRNet adaptation maintains high-resolution inputs, making it particularly well-suited for tasks requiring detailed image analysis, such as medical imaging and remote sensing [10].

The remainder of this paper is organized as follows: Section II discusses the literature review, highlighting the advancements and limitations of current models. Section III details the development of our customized HRNet model, elaborating on the specific modifications and their intended benefits. Section IV presents the outcomes of our model, showcasing its performance improvements over existing models. Through this structured approach, we aim to contribute significantly to the semantic segmentation domain, addressing critical challenges and setting a new benchmark for future research.

LITERATURE SURVEY

Du Jiang et al [10] presented a method for multi-task semantic segmentation for use in an environment as complex as an indoor one. This method is based on an upgraded version of the Faster-RCNN algorithm and is used for joint target identification using RGB-D image information. This model is capable of realising numerous visual tasks concurrently, including semantic segmentation of an interior scene, target categorization, and detection of multiple objects. Here, the way RGB photos and depth photos are put together is made better by the effect that uneven lighting in the surroundings has. It not only gives more knowledge about the features of the fusion picture, but it also makes model training more effective.

Laigang Zhang et al [11] proposed an unsupervised co-segmentation technique, and it has the capability of being used on images that include many foreground objects at the same time while the background is undergoing significant change. In order to do semantic extraction, the colour edge image in RGB space is first retrieved. This method offers a time-saving approach to distinguishing between the foreground and the background by iteratively modelling the appearance pattern of pixel and area populations. It is possible to strengthen the coherence of the image's background and foreground model by making advantage of the correlation that exists between the different picture sections and the interior of the image.

Seonkyeong Seong et al [12] extracted of buildings from aerial photos was performed using the csAG-HRNet approach, which involved the utilization of HRNet-v2 in combination with channel and spatial attention gates. HRNet-v2 encompasses transition and fusion mechanisms structured around subnetworks tailored to handle various resolutions. Within this framework, the network harnessed channel attention gates to allocate weights based on the relative significance of each channel, while spatial attention gates assigned weights based on the relative value of each pixel location within the entire channel. The architecture of csAG-HRNet introduced csAG modules, consisting of a channel attention gate and a spatial attention gate, which were applied to each subnetwork within the stage and fusion modules of the HRNet-v2 network.

Hatem Ibrahem et al [13] utilized designs that leverage depth-wise separable convolutions. In this study, we present two distinct networks, namely DTS-Net and DTS-Net-Lite, which are designed for real-time semantic segmentation. The DTS-Net utilizes a deep network architecture, employing Xception as the feature extractor. On the other hand, the DTS-Net-Lite adopts a lightweight network architecture, using MobileNetV2 as the feature extractor. Furthermore, we investigate the challenge of combined semantic segmentation and depth estimation and provide evidence that the proposed methodology is capable of effectively executing both tasks concurrently, surpassing current state-of-the-art (SOTA) approaches.

Ilias Papadeas et al [14] provided a complete review of the current state-of-the-art approaches for semantic picture segmentation. Specifically, it focuses on deep-learning techniques that are designed to function in real time, with the goal of efficiently supporting autonomous driving scenarios. In order to achieve this objective, the provided overview places significant importance on the presentation of approaches that enable the reduction of inference time. Additionally, an analysis of the current methods is conducted by considering their end-to-end functionality, and a comparative study is conducted using a standardized evaluation framework.

Rubén Usamentiaga et al [15] examined the current advanced deep-learning techniques used in object recognition and semantic segmentation within the domain of automated surface inspection in the metal industry. Images obtained inside the industrial setting are subject to environmental factors such as noise, dust, and vibrations, hence presenting an extra set of challenges. Furthermore, the process of industrial inspection necessitates not just precision, but also resilience and efficiency.

L. Minh Dang et al [16] proposed an automated framework for segmenting sewage defects at the pixel level, using the DeepLabV3+ model. The system aims to accurately identify the kind, location, geometric characteristics, and severity of the detected defects. A comprehensive evaluation was conducted to assess the effects of different backbones and pre-processing techniques on the performance of the model. Furthermore, the comparative analysis included four contemporary segmentation models, namely U-Net, SegNet, PSPNet, and FCN, in order to showcase the improved performance of the proposed model.

Yanling Zou et al [17] proposed a novel framework for the training and evaluation process that encompasses a representative metropolitan setting in central Europe, namely Munzingen, Freiburg, located in the state of Baden-Württemberg, Germany. The dataset used in this study comprises over 390 million dense points obtained via the use of Mobile Laser Scanning (MLS). This dataset exhibits a much greater number of sample points compared to current datasets and encompasses relevant item categories that are specifically relevant to applications in the domains of smart cities and urban planning. In this study, we conduct a comprehensive evaluation of the Deep Neural Network (DNN) using our dataset. We explore many significant issues from several perspectives, including data preparation procedures, the potential benefits of including color information, and the issue of uneven class distribution often seen in real-world scenarios.

Yuzhu Ji et al [18]provided a more comprehensive understanding of the underlying designs of important components and frameworks in CNN-based encoder-decoders for applications requiring dense prediction at the pixel level. This viewpoint is expressed via the perspective of fundamental architectural forms. Additionally, the authors stress the use of deep encoder-decoder models for SOD and provide a complete empirical analysis of baseline encoder-decoder models in terms of different encoder backbones, loss functions, training batch sizes, and attention structures. Additionally, research is being conducted on sophisticated encoder-decoder techniques derived from semantic segmentation as well as deep CNN-based SOD algorithms. It has been shown that fresh baseline models can outperform state-of-the-art models.

Chang Sun et al [19] proposed the method known as a "guided SSD" (Mask-SSD), which has two branches: a segmentation branch and a detection branch. A feature-fusion module was made so that the recognition branch could use environmental data to make high-resolution feature maps. Most of the segmentation branch was made with arous convolution to give the detecting branch more context information. The data that went into the segmentation branch also went into the detection branch, and the features made by the segmentation branch were combined with the features made by the detection branch to classify and find the target items. Also, segmentation features were used to make the mask, which was then used to tell the detection branch to look for things in the "foreground" areas.

Shibao Li et al [20] developed an object recognition to semantic segmentation (O2S) tool that can be plugged in and used right away. This will allow object detectors to fully use the training set, including intermediate feature maps made by object detectors, while also making accurate predictions about semantic masks. Also, the authors provide an object recognition training set-based box-level weakly guided probability gap adaptive (PGA) method that lets O2S learn semantic masks.

Márton Szemenyei et al [21] presented two distinct neural network topologies that guarantee low-cost inference while boosting accuracy by using the elements of the environment for semantic segmentation and object recognition. These designs are designed to improve semantic segmentation and object recognition. These models are able to learn from a small number of manually annotated photos because to a technique called synthetic transfer learning.

The shortcomings of the existing models include limited performance in handling complex environments, inefficiencies in real-time processing, a lack of specialized models for pixel-level segmentation in specific applications, difficulties in effectively processing large and diverse datasets for urban planning and smart cities, and a need for advanced encoder-decoder architectures for more accurate salient object detection. These models often lack seamless integration of object detection and semantic segmentation, and they underutilize object recognition training sets for segmentation tasks.The solution lies in the development of a model that can effectively handle the complexities of outdoor environments, while also performing high-quality semantic segmentation and object detection.

PROPOSED MODEL

This section elucidates the architectural design and foundational principles of the proposed High-Resolution Network (HRNet) tailored specifically for the task of semantic segmentation. Central to HRNet's innovation is its ability to process inputs at high resolutions throughout its architecture, thereby retaining intricate image features essential for precise segmentation. Unlike conventional deep learning models, which often compromise on input resolution to alleviate computational burdens, HRNet adopts a multi-resolution representation approach. This methodology permits the concurrent processing of the input image across various spatial resolutions through distinct parallel streams. These streams are intricately merged at key points within the network, enabling an integrative synthesis of detailed attributes and broader contextual cues.

The HRNet model is ingeniously designed to efficiently manage high-resolution inputs, ensuring exceptional accuracy levels. Its unique strategy involves a multi-resolution representation of the input image, sidestepping the common practice among traditional deep learning frameworks of downscaling the image to a lower resolution. This network is structured with several parallel branches, each dedicated to processing the image at a distinct spatial resolution. As these branches converge across various stages within the network, they facilitate a dynamic interplay between minutely detailed observations and the overarching contextual landscape. The modular architecture of HRNet, underscored by its high-resolution module (HRM), stands as a testament to its adaptability, allowing for seamless adjustments in response to varying input sizes or computational demands. Each HRM branch, operating in parallel, generates a feature map, which is then intricately fused with outputs from other branches through a meticulously designed high-to-low resolution fusion process. This process is visually represented in Figure 1, showcasing the architectural layout of HRNet.

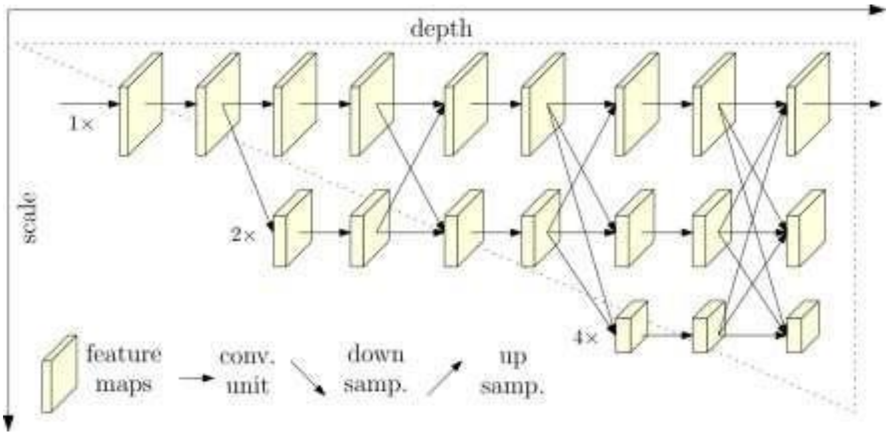


Figure 1: The HRNet Architecture

A deep network may be created by stacking the HRM many times, which enables the model to capture the input image's more complicated characteristics and hierarchical representations. A segmentation map showing the position and boundaries of the objects in the input picture is the network's final output. In a number of computer vision tasks, such as object identification, human position estimation, and semantic segmentation, HRNet has demonstrated cutting-edge performance. The architecture's ability to handle high-resolution inputs efficiently, capture fine-grained details and coarse contextual information, and produce accurate segmentations has made it a popular choice in the computer vision community.

3.1 ProposedHRNet model

The proposed model is designed to perform image segmentation architecture, and it is based on High-Resolution Network (HRNet). This design makes use of the complementary strengths of two components, known as the "feature_block" and the "Extended_feature_block." To take use of the benefits of multi-scale feature extraction, these building blocks are combined, which ensures that the network can efficiently collect both fine-grained features and contextual information from the input data. This network exhibits a capacity to improve the accuracy of image segmentation since it uses these building blocks and combines data across multiple scales. As a result, it is a flexible and effective tool that can be used for a broad variety of computer vision applications including video surveillance.

*Mathematical Formulation of HRNet:*Consider an input image I, which is fed into the HRNet architecture. The network utilizes multiple parallel branches $B_1, B_2, \ldots, B_n$, each responsible for processing I at distinct spatial resolutions. Let $f_{B_i}(I)$ denote the feature map produced by branch B_i. The HRNet's core mechanism, the HighResolution Module (HRM), orchestrates the fusion of these feature maps across resolutions.

Mathematically, the fusion process at a given stage can be represented as:

$$F_{\text{merged}} = \sum_{i=1}^{n} \alpha_i \cdot \text{Upsample}\ (f_{B_i}(I)) \qquad (1)$$

where F_{merged} is the merged feature map, α_i are learnable weights, and Upsample $(\cdot)$ denotes the upsampling operation to match the dimensions across different resolutions.

A. Feature Block

The feature block (Fb) is pivotal in extracting meaningful features from I, constituted of two convolutional layers complemented by batch normalization and ReLU activation functions. For an input tensor x, the operations within Fb can be expressed as:

$$y_{\text{output}} = \text{ReLU}\ (\text{BatchNorm}\ (\text{Conv}_{3x3}\ (\text{ReLU}\ (\text{BatchNorm}\ (\text{Conv}_{1x1}\ (x)))))) \qquad (2)$$

This layered configuration is instrumental in distilling local features from x, thereby amplifying the network's competence in capturing nuanced details pivotal for accurate segmentation.

Pseudocode: feature_block (Fb)

```
x: the input tensor.
W: weight
b: Bias
y1=Conv1x1(x, W1, b1)
y2=BatchNormalization(y1)
y3=ReLU(y2)
y4=Conv3x3(y3, W2, b2)
y5=BatchNormalization(y4)
y6=ReLU(y5)
$y$output=$y$6
```

The procedural execution within the Feature Block unfolds as follows:

1. Apply a 1 × 1 convolution to x using weights W_1 and bias b_1, resulting in y_1.
2. Perform batch normalization on y_1, yielding y_2.
3. Introduce non-linearity through the ReLU function on y_2, obtaining y_3.
4. Conduct a 3 × 3 convolution on y_3 with weights W_2 and bias b_2, producing y_4.
5. Normalize y_4 using batch normalization, leading to y_5.
6. Apply the ReLU function to y_5 to generate the final output y_{output}.

This sequence initiates with a 3 × 3 convolutional layer equipped with 32 filters acting upon x, followed by batch normalization to standardize the activations to a zero mean and unit variance. This standardization is crucial for stabilizing the training process. Subsequently, a ReLU activation function infuses non-linearity into the processed data, vital for capturing complex patterns within the input tensor. The iterative application of another 3 × 3 convolutional layer, coupled with batch normalization and ReLU activation, further refines the feature extraction process, culminating in the generation of y_{output}.

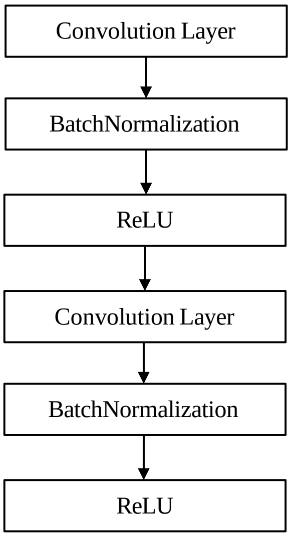


Figure 2: Feature Block

Figure 2 encapsulates the operational paradigm of the Feature Block, highlighting its role as a cornerstone in the HRNet architecture for discerning and amplifying critical local features indispensable for accurate semantic segmentation.

B. Extended Feature Extraction

The Extended Feature Extraction phase represents a pivotal enhancement within the proposed High-Resolution Network (HRNet) architecture, meticulously designed to refine the feature extraction process through advanced convolutional techniques and activation functions. This phase is characterized by a sequence of operations that not only amplify the detail and quality of extracted features but also address key challenges inherent in deep neural network training, such as the vanishing gradient problem.

Pseudocode: Extend Feature Extraction

```
y1=Conv1x1(x, W1, b1)
y2=BatchNormalization(y1)
y3=SeLU(y2)
y4=Conv3x3(y3, W2, b2)
y5=BatchNormalization(y4)
y6=SeLU(y5)
y7= SeparableConv2D(y6, W3, b3)
y8=BatchNormalization(y7)
y9=SeLU(y8)
youtput=y9
```

The Extended Feature Extraction initiates with the application of a $2D$ convolutional layer to the input tensor x, equipped with 32 filters of kernel size 1×1. This operation can be mathematically expressed as:

$$y_1 = \mathrm{Conv}_{1\times 1}\,(x, W_1, b_1)$$

Following this, batch normalization is applied to y_1, standardizing its activations to a normalized distribution, which is essential for enhancing the stability of the training dynamics. This step can be represented as:

$$y_2 = \mathrm{BatchNormalization}\,(y_1)$$

The Scaled Exponential Linear Unit (SELU) activation function is then employed to introduce nonlinearity into the process while promoting self-normalization of the activations. The SELU operation on y_2 is denoted as:

$$y_3 = \mathrm{SeLU}\,(y_2)$$

Proceeding further, another $2D$ convolutional layer with 32 filters of a larger kernel size 3×3 is applied to y_3, followed by batch normalization and SELU activation, akin to the initial steps. This enhances the capability to capture more complex patterns within the data:

$$\begin{aligned} y_4 &= \mathrm{Conv}_{3\times3}\,(y_3, W_2, b_2) \\ y_5 &= \mathrm{BatchNormalization}\,(y_4) \\ y_6 &= \mathrm{SeLU}\,(y_5) \end{aligned} \tag{3}$$

The subsequent introduction of a separable convolutional layer, consisting of 96 filters with a kernel size of 3×3, further delineates the feature extraction process. The separable convolution, a composite of depthwise and pointwise convolutions, significantly reduces the parameter count, enhancing the network's computational efficiency:

$$\begin{aligned} y_7 &= \mathrm{SeparableConv2D}\,(y_6, W_3, b_3) \\ y_8 &= \mathrm{BatchNormalization}\,(y_7) \\ y_9 &= \mathrm{SeLU}\,(y_8) \end{aligned} \tag{4}$$

The output of this extended feature extraction phase, $y_{\text{output}} = y_9$, encapsulates a rich, hierarchical representation of features, optimally prepared for subsequent segmentation tasks.

Advantages of SELU and SeparableConv2D

The utilization of the SELU activation function within this context serves to mitigate the vanishing gradient issue by fostering an environment where self-normalization of neurons occurs naturally. This leads to more stable and consistent training phases, thereby circumventing common pitfalls associated with deep learning models. The SELU function is defined as:

$$\mathrm{SELU}\,(x) = \lambda \begin{cases} x & \text{if } x > 0 \\ \alpha\left(e^{x} - 1\right) & \text{if } x \le 0 \end{cases} \tag{5}$$

where λ and α are predefined scaling parameters that ensure the activations self-normalize, promoting a healthy gradient flow.

Conversely, the SeparableConv2D layer optimizes computational resource usage by deconstructing the convolutional process into depthwise followed by pointwise operations, thus curtailing the overall parameter volume required. This decomposition not only conserves memory but also accelerates the computation, making the architecture more feasible for deployment in resource-constrained environments.

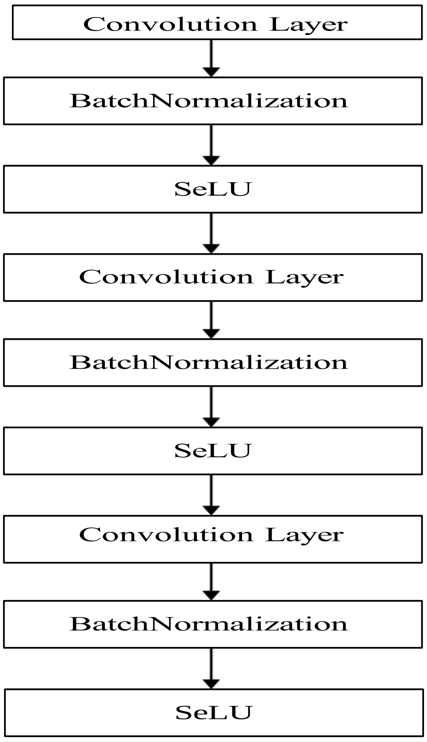


Figure 3: Extended Feature Block

In summary, the Extended Feature Extraction stage of the proposed HRNet model is a testament to the architectural ingenuity aimed at extracting highly detailed and hierarchical feature representations. Through a judicious combination of advanced convolutional techniques and the strategic implementation of SELU and SeparableConv2D, this phase significantly elevates the model's performance, paving the way for achieving state-of-the-art results in semantic segmentation tasks.

C. Proposed HRNet model

The architecture combines the features learned at different resolutions or scales to enhance the model's ability to perform effective segmentation.

1. **Feature Extraction:** The model begins by performing feature extraction using a series of convolutional layers, batch normalization, and ReLU activations. These layers aim to capture various levels of image features.
2. **Residual Connections:** Residual connections are introduced to preserve and propagate low-level image details throughout the network. These connections help mitigate the vanishing gradient problem and allow the network to learn fine-grained details.
3. **Multi-Scale Processing:** The network processes information at multiple scales. It applies different strategies to handle features at different resolutions. This helps in handling both local and global context efficiently.
4. **Combining Scales:** The code combines features from multiple scales through upsampling, 1x1 convolutions, and element-wise addition. This fusion of features from different scales ensures that the model can make accurate segmentation predictions by considering features from both fine and coarse levels of detail.
5. **Final Segmentation Output:** After combining features from different scales, the code applies a final convolutional layer to generate the segmentation output. This layer produces a pixel-wise prediction for the segmentation task.

UConv: Upsampling layer with Convolution Layer

UConvBR: Upsampling layer with Convolution Layer, followed by batch normalization and ReLu

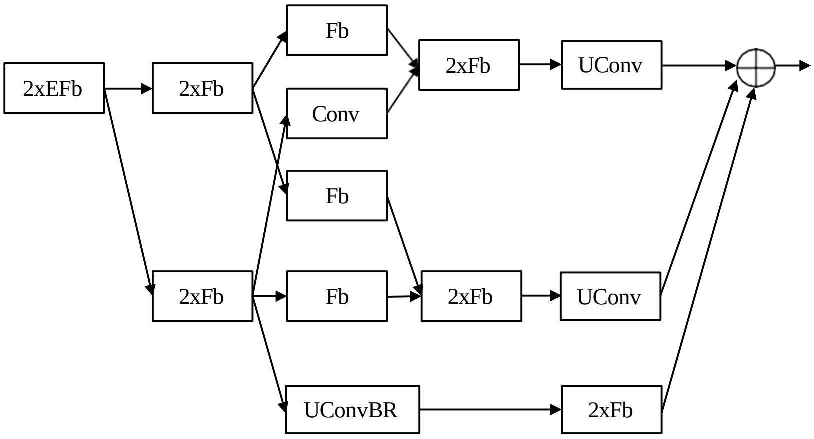


Figure 4: Proposed HRNet model

The mathematical analysis involves understanding the impact of various operations:

- **Convolutional Layers:** Convolution operations apply linear transformations to the input data. The kernel size and the number of filters determine the receptive field and the complexity of the features learned.
- **Separable convolution Layers:** Separable convolution layers are a specific variant of convolutional neural network (CNN) layers that aim to decrease computational complexity while still being able to capture significant spatial hierarchies in input. Separable convolution, in contrast to typical convolutional layers, divides the operation into two distinct steps: depthwise convolution and pointwise convolution. Instead of performing a full convolution with a kernel across all input channels, separable convolution breaks down the process into these two separate operations. During the depthwise convolution process, each input channel undergoes convolution separately using its own distinct set of filters. Next, the pointwise convolution is performed, which involves using a 1x1 convolution to merge the output channels obtained from the depthwise

convolution. This segregation greatly decreases the quantity of parameters and computations, resulting in a more efficient network structure.

- **Batch Normalization:** Batch normalization normalizes the activations, reducing internal covariate shift and speeding up convergence.
- **ReLU Activation:** The ReLU activation function introduces non-linearity into the model, allowing it to capture complex patterns.

$$ReLU(x) = max(0, x) \qquad (6)$$

- **SeLu Activation:** The Scaled Exponential Linear Unit (SELU) is an activation function designed to address certain challenges associated with other activation functions, such as vanishing and exploding gradients. One distinctive feature of SELU is its ability to induce self-normalizing properties in neural networks.

$$SELU(x) = \lambda * \{ \begin{array}{cc} x & if\ x > 0 \\ \alpha * \exp(x) - 1 & if\ x < 0 \end{array} \qquad (7)$$

λ and α are constants

- **Upsampling:**Upsampling is used to increase the spatial resolution of feature maps. This operation is often paired with 1x1 convolutions to adjust the number of channels.
- **Element-wise Addition:** The addition operation combines feature maps, enabling information flow between different scales of the network.
- **Residual Connections:** Residual connections allow the network to learn the residual (difference) between the current and previous layer's output. This helps in preventing vanishing gradients and retaining low-level details.

The combination of these operations and the multi-scale processing strategy contribute to the model's ability to perform effective segmentation. By processing features at different resolutions and preserving fine details, the model can segment objects and regions accurately. Additionally, the incorporation of residual connections and feature fusion mechanisms ensures that the model can handle complex structures and textures in the input images effectively. The proposed model is designed to balance between local and global context, making it highly suitable for image segmentation tasks, particularly for high-resolution images.

The goal is to enhance the model's ability to capture and utilize multi-resolution information. This is performed by using the following:

- The proposed model increases the number of output channels in the third convolution (1x1) from 64 to 96.
- It incorporates upsampling and concatenation of features from different branches, providing the network with a mechanism to handle multi-resolution information.
- The model has increased number of channels in the Extended Feature extraction module and the use of concatenation in the final output, leads to improved feature representation and better information fusion.

A. Loss function

In the proposed model, HRNet is added with DICE (Dice Similarity Coefficient) loss function. DICE loss function is a measure of the overlap between two sets. It is commonly used as a loss function for segmentation tasks because it penalizes false positives and false negatives equally, unlike other loss functions such as cross-entropy, which tend to penalize false positives more. Using the DICE loss function with HRNet can improve the network's performance in semantic segmentation tasks by encouraging the network to produce more accurate segmentations. The following definition applies to the DICE loss function:

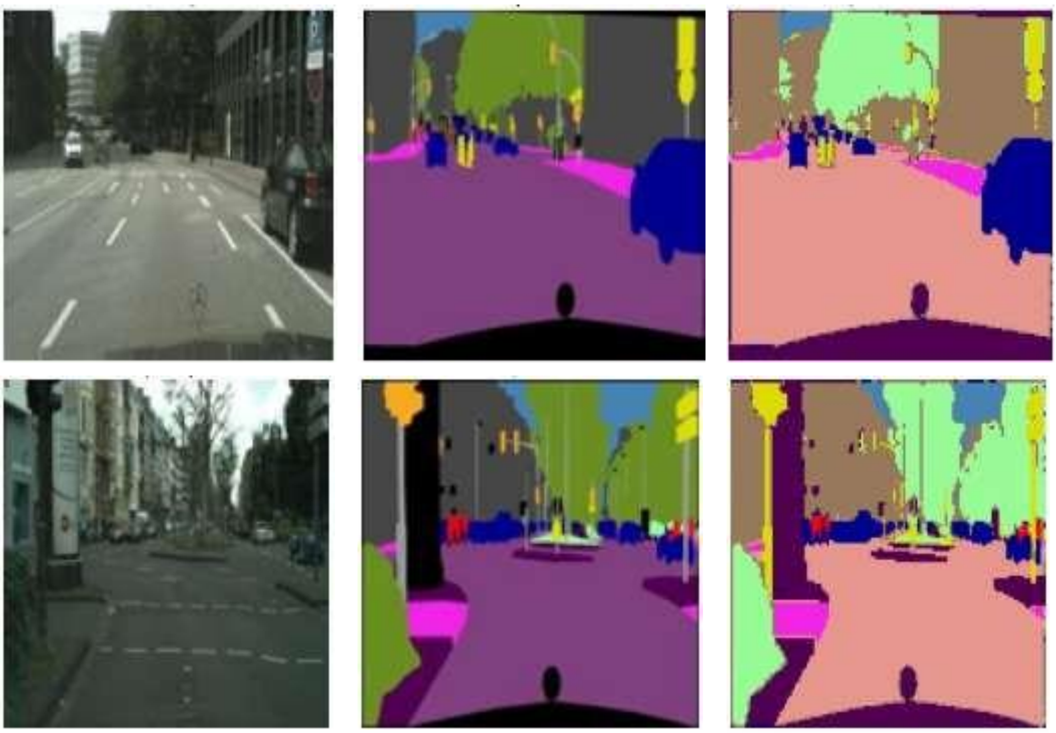
$$DICE\ Loss = 1 - (2 * Intersection\ /\ (Union + Intersection)) \qquad (8)$$

where Union is the total pixels present in the predicted as well as ground truth masks, and Intersection is the number of pixels that are properly identified as being members of the target class.

HRNet with DICE loss function may enhance semantic segmentation performance by successfully handling the class imbalance issue in training data. The class imbalance issue happens when the number of pixels in the training set belonging to one class greatly outnumbers the pixels belonging to other classes. In such cases, the model can be biased towards the majority class, resulting in poor segmentation performance for the minority classes. The DICE loss function has been shown to be effective in handling class imbalance by weighting each class inversely proportional to its frequency in the training set. By incorporating the DICE loss function into the HRNet architecture, the model can be trained to segment each class more accurately, including the minority classes. This can result in improved segmentation performance for the entire image. Furthermore, the HRNet architecture's multi-resolution representation learning capabilities, combined with the DICE loss function, can help capture more fine-grained details in the input image, resulting in better segmentation performance.

SIMULATION RESULTS

This section discusses the simulation results of object segmentation by using proposed HRNet architecture. The Cityscapes data, which consists of labelled videos collected from vehicles while they were being driven in Germany, is used in this study. The dataset includes still photos that were taken from the original films; moreover, the semantic segmentation labels are shown in images alongside the original image. The sample dataset images are depicted in Figure 5.



(a) Sample Image (b) Sample mask (c) Smaple Encoded mask

Figure 5: Input images from Cityscapes dataset

This dataset contains 500 validation image files in addition to the 2975 training image files that are included. Each picture file is 256 by 512 pixels in size, and each file is a composite consisting of the annotated image (the result of semantic segmentation) on the right half of the image, together with the original photo on the left half of the image.

In this object segmentation, validation accuracy is used as one of the performance metrics which measure of how well a proposed HRNet model can accurately identify objects in images. It is commonly calculated by comparing the model's predicted object segmentation masks to the validation dataset's ground truth segmentation masks. Similarly, the validation loss measures how effectively a model generalises to new, previously unknown data. It is derived by subtracting the predicted object segmentation masks generated by the model from the ground truth segmentation masks in the validation dataset. The validation loss is used to measure how well the model works during training and figure out if it is overfitting or underfitting. Figure 6 shows that the proposedHRNet mode is no longer valid.

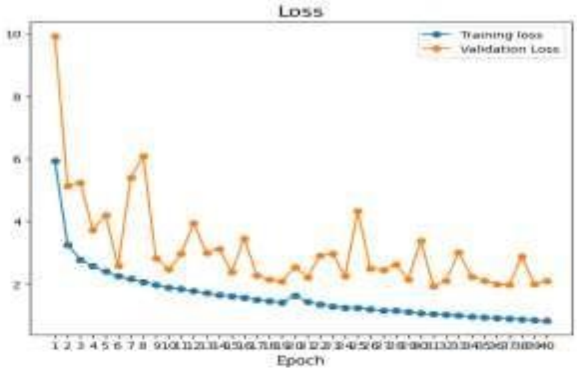


Figure 6: Training and Validation loss of the proposed model

The second performance statistic is the intersection over Union (IoU). The IoU is a popular statistic for evaluating object segmentation tasks. It calculates the amount of overlap between the expected and actual object segmentation. The ratio of the intersection of the anticipated as well as ground truth segmentation masks to the union of the identical masks is used to compute IoU. It may be stated mathematically as:

$$IoU = \frac{intersection\ of\ predicted\ and\ ground\ truth\ masks}{union\ of\ predicted\ and\ ground\ truth\ masks} \tag{9}$$

A high IoU value indicates a good segmentation performance, as it means that the predicted segmentation closely matches the ground truth segmentation. Conversely, a low IoU value suggests that the predicted segmentation is not accurate. Figure 5 shows the proposed HRNet model's validation and training IoU graph.

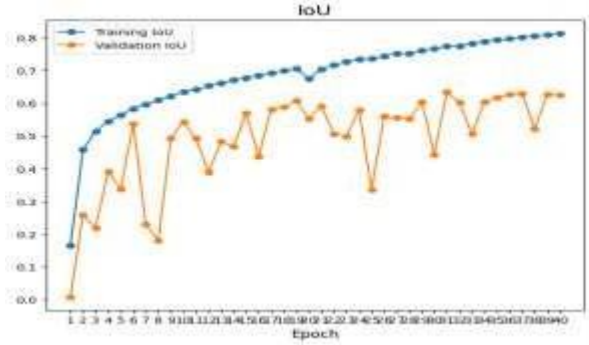


Figure 7: Training and Validation IoU of the proposed model

To assess the effectiveness of object segmentation algorithms, IoU is often used in conjunction with other metrics like as accuracy, recall, and F1 score. It is also used in training neural networks for object segmentation tasks, where the goal is to maximize IoU during training.

Validation accuracy is crucial since it helps to assess the model's performance and offers an estimate of how well the model will perform on fresh, previously unseen data. High validation accuracy indicates that the model is accurately segmenting objects in images, while low accuracy suggests that the model may be incorrectly segmenting objects or failing to identify them altogether.The proposed model's validation accuracy as well as IoU are compared to comparable network models [29].

Table 1: Proposed model evaluation

Parameter	**Value**
mIoU	63.43%
Validation accuracy	85.8%

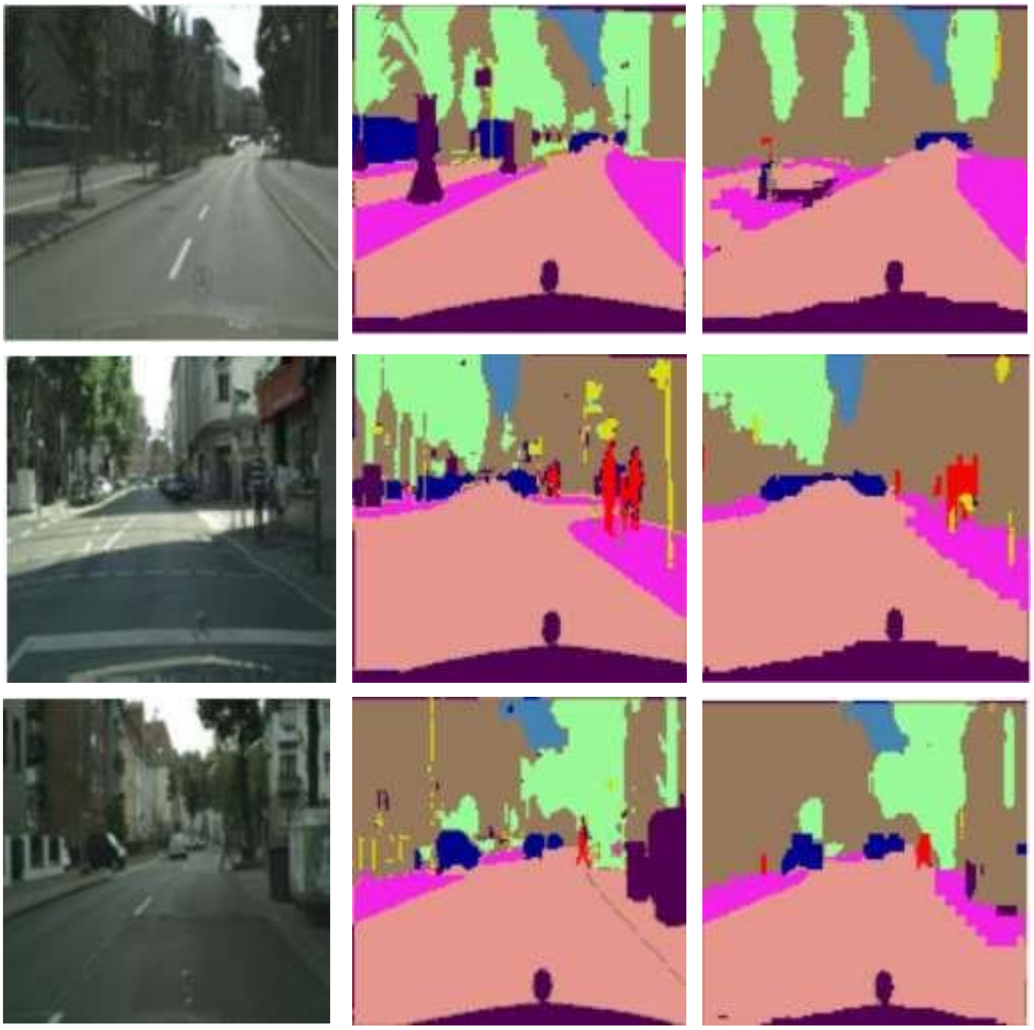
Table 2 shows the validation accuracy comparison table.

Table 2: Comparative analysis

Model	**Validation Accuracy (%)**
FCN+ Transfer learning [21]	32.21
Unet [22]	83.03
C-Unet [23]	83.35
Conventional HRNet [24]	82.41

Proposed HRNet	**85.8**

Figure 8 shows the segmentation results of the proposed model.



(a) Input Image (b) Encoded Mask (c) Predicated Image

Figure 8: Segmentation Results of Proposed HRNet Model

The comprehensive analysis presented in Table 2 delineates a comparative study focused on the validation accuracy achieved by various semantic segmentation models, including the proposed High-Resolution Network (HRNet) model. This table encapsulates a critical evaluation of the performance metrics obtained across a spectrum of state-of-the-art models, underpinning the advancements facilitated by the proposed HRNet architecture in the domain of semantic segmentation.

Commencing with the FCN augmented with Transfer Learning [22], it is observed that this model achieves a validation accuracy of 32.21%. This relatively modest performance underscores the challenges inherent in adapting pre-trained networks to the nuanced requirements of semantic segmentation tasks. In contrast, the Unet [23] model demonstrates a significant improvement, securing a validation accuracy of 83.03%. The incremental enhancement is further evident in the C-Unet [24] model, which slightly surpasses its predecessor with an accuracy of 83.35%, highlighting the benefits of context incorporation and advanced feature extraction strategies.

The conventional HRNet [25] model, renowned for its high-resolution processing capabilities, achieves a validation accuracy of 82.41%. While this figure represents a commendable achievement, especially considering the model's intricate architecture designed to preserve high-resolution details throughout the segmentation process, it falls short when juxtaposed against the enhanced methodologies employed in the subsequent models[26][27].

Pivotal to this comparative analysis is the performance of the proposed HRNet model, which stands at a validation accuracy of 85.8%. This remarkable achievement not only surpasses the aforementioned models but also sets a new benchmark within the realm of semantic segmentation. The proposed HRNet model's superior accuracy can be attributed to its innovative modifications, including optimized feature blocks and extended feature extraction phases that collectively enhance the model's ability to discern and segment complex images with heightened precision[28].

The validation accuracy comparison elucidated in Table 2 unequivocally underscores the efficacy of the proposed HRNet model in addressing and surmounting the challenges posed by semantic segmentation tasks. By achieving a notable accuracy of 85.8%, the proposed model heralds a significant leap forward, promising to propel the boundaries of what is achievable in the field of computer vision and semantic image segmentation. This analysis not only validates the theoretical and architectural merits of the proposed HRNet but also accentuates its practical applicability and superiority in generating highly accurate segmentation results.

4.1 Limitations of the Study : Despite the significant advancements demonstrated by the proposed High-Resolution Network (HRNet) model in semantic segmentation, certain limitations merit consideration. Firstly, while the model exhibits superior validation accuracy, its performance across diverse and more complex datasets, particularly those with highly irregular or small objects, has not been extensively explored. The intricacies of such datasets may pose challenges that could affect the generalizability of the model's effectiveness.

Secondly, the computational efficiency, although enhanced compared to conventional HRNet models, remains a concern, especially when deploying the model in real-time applications or on devices with limited processing capabilities. The balance between high-resolution processing and computational demand is crucial, and further optimization may be necessary to ensure broader applicability.

Moreover, the study predominantly focuses on urban scene segmentation, potentially limiting the insights into the model’s performance in varied domains such as medical imaging, aerial imagery, and natural scene understanding. Each of these domains presents unique challenges that could influence the model's adaptability and performance.

4.2 Future Directions: Addressing the limitations outlined above opens several avenues for future research. Expanding the model’s evaluation to encompass a wider array of complex datasets, including those with small or irregularly shaped objects, would provide deeper insights into its generalizability and effectiveness across different segmentation challenges.

Further, there is a pressing need for optimization strategies that specifically target computational efficiency. Research into model pruning, quantization, and the development of lightweight versions of the proposed HRNet could facilitate its deployment in real-time applications and on edge devices, thereby widening its scope of applicability.

Investigations into transfer learning and domain adaptation strategies present another promising direction. By adapting the proposed HRNet model to various domains beyond urban scenes, such as medical imaging or aerial photography, researchers can explore the model's versatility and performance in segmenting diverse image types.

Lastly, the integration of attention mechanisms and advanced feature fusion techniques could enhance the model’s ability to focus on relevant image regions and improve segmentation accuracy further. Exploring the synergy between these techniques and the proposed HRNet’s architecture may yield significant improvements in semantic segmentation tasks.

In conclusion, while the proposed HRNet model marks a significant advancement in semantic segmentation, exploring these future directions could address current limitations and unlock new possibilities in the field of computer vision. Through continuous innovation and exploration, the potential of deep learning models like HRNet in semantic segmentation can be fully realized, paving the way for more accurate, efficient, and versatile segmentation solutions.

CONCLUSION

HRNet is highly effective for semantic segmentation tasks. The key features of the model include features, including multi-scale feature fusion, high-resolution feature extraction, multi-pathway aggregation, and efficient training. HRNet's multi-scale feature fusion allows it to capture features at multiple levels of abstraction, while maintaining high-resolution features throughout the network. This enables the network to capture fine-grained details and accurately segment objects of varying sizes. HRNet can collect both low-level as well as high-level information because to multi-pathway aggregation, which is crucial for effectively segmenting complicated objects. HRNet's efficient training approach allows it to converge faster and achieve better performance. It integrates the process of upsampling and combining features from various branches, which equips the network with the capability to manage multi-resolution information effectively. The Extended Feature Extraction module has been enhanced with SeLu and separable convolution layers, and the incorporation of concatenation in the final output results in improved feature representation and more effective information fusion. The proposed HRNet is a very effective framework for semantic segmentation tasks, as the proposed model obtained a validation accuracy of 85.8% and mean IoU of 63.43%, outperforming the other models.

REFERENCES

[1] Pu, Huayan, Jun Luo, Gang Wang, Tao Huang, and Hongliang Liu. "Visual SLAM Integration with Semantic Segmentation and Deep Learning: A Review." *IEEE Sensors Journal* (2023).

[2] Károly, Artúr István, Sebestyén Tirczka, Huijun Gao, Imre J. Rudas, and Péter Galambos. "Increasing the Robustness of Deep Learning Models for Object Segmentation: A Framework for Blending Automatically Annotated Real and Synthetic Data." *IEEE Transactions on Cybernetics* (2023).

[3] Dodla. Likhith Reddy, &Dr. D Prathyusha Reddi. (2017). Texture Image Segmentation Based on threshold Techniques. International Journal of Computer Engineering in Research Trends, 4(3), 69–75.

[4] Jaiswal, Sushma, and M. K. Pandey. "A review on image segmentation." *Rising Threats in Expert Applications and Solutions: Proceedings of FICR-TEAS 2020* (2020): 233-240.

[5] Minaee, Shervin, Yuri Boykov, Fatih Porikli, Antonio Plaza, Nasser Kehtarnavaz, and Demetri Terzopoulos. "Image segmentation using deep learning: A survey." *IEEE transactions on pattern analysis and machine intelligence* 44, no. 7 (2021): 3523-3542.

[6] Bayoudh, Khaled, Raja Knani, Fayçal Hamdaoui, and Abdellatif Mtibaa. "A survey on deep multimodal learning for computer vision: advances, trends, applications, and datasets." *The Visual Computer* (2021): 1-32.

[7] Gao, Mingqi, Feng Zheng, James JQ Yu, Caifeng Shan, Guiguang Ding, and Jungong Han. "Deep learning for video object segmentation: a review." *Artificial Intelligence Review* 56, no. 1 (2023): 457-531.

[8] Ji, Ankang, Alvin Wei Ze Chew, Xiaolong Xue, and Limao Zhang. "An encoder-decoder deep learning method for multi-class object segmentation from 3D tunnel point clouds." *Automation in Construction* 137 (2022): 104187.

[9] Dijkstra, Klaas, Jaap van de Loosdrecht, Waatze A. Atsma, Lambert RB Schomaker, and Marco A. Wiering. "CentroidNetV2: A hybrid deep neural network for small-object segmentation and counting." *Neurocomputing* 423 (2021): 490-505.

[10] Jiang, Du, Gongfa Li, Chong Tan, Li Huang, Ying Sun, and Jianyi Kong. "Semantic segmentation for multiscale target based on object recognition using the improved Faster-RCNN model." *Future Generation Computer Systems* 123 (2021): 94-104.

[11] Zhang, Laigang, Zhou Sheng, Yibin Li, Qun Sun, Ying Zhao, and Deying Feng. "Image object detection and semantic segmentation based on convolutional neural network." *Neural Computing and Applications* 32 (2020): 1949-1958.

[12] Seong, Seonkyeong, and Jaewan Choi. "Semantic segmentation of urban buildings using a high-resolution network (HRNet) with channel and spatial attention gates." *Remote Sensing* 13, no. 16 (2021): 3087.

[13] Ibrahem, Hatem, Ahmed Salem, and Hyun-Soo Kang. "DTS-Net: Depth-to-Space Networks for Fast and Accurate Semantic Object Segmentation." *Sensors* 22, no. 1 (2022): 337.

[14] Papadeas, Ilias, Lazaros Tsochatzidis, Angelos Amanatiadis, and Ioannis Pratikakis. "Real-time semantic image segmentation with deep learning for autonomous driving: A survey." *Applied Sciences* 11, no. 19 (2021): 8802.

[15] Usamentiaga, Rubén, Dario G. Lema, Oscar D. Pedrayes, and Daniel F. Garcia. "Automated surface defect detection in metals: A comparative review of object detection and semantic segmentation using deep learning." *IEEE Transactions on Industry Applications* 58, no. 3 (2022): 4203-4213.

[16] Dang, L. Minh, Hanxiang Wang, Yanfen Li, Le Quan Nguyen, Tan N. Nguyen, Hyoung-Kyu Song, and Hyeonjoon Moon. "Lightweight pixel-level semantic segmentation and analysis for sewer defects using deep learning." *Construction and Building Materials* 371 (2023): 130792.

[17] Zou, Yanling, Holger Weinacker, and Barbara Koch. "Towards urban scene semantic segmentation with deep learning from LiDAR point clouds: a case study in Baden-Württemberg, Germany." *Remote Sensing* 13, no. 16 (2021): 3220.

[18] Ji, Yuzhu, Haijun Zhang, Zhao Zhang, and Ming Liu. "CNN-based encoder-decoder networks for salient object detection: A comprehensive review and recent advances." *Information Sciences* 546 (2021): 835-857.

[19] Sun, Chang, Yibo Ai, Sheng Wang, and Weidong Zhang. "Mask-guided SSD for small-object detection." *Applied Intelligence* 51 (2021): 3311-3322.

[20] Li, Shibao, Yixuan Liu, Yunwu Zhang, Yi Luo, and Jianhang Liu. "Adaptive Generation of Weakly Supervised Semantic Segmentation for Object Detection." *Neural Processing Letters* (2022): 1-14.

[21] Szemenyei, Márton, and Vladimir Estivill-Castro. "Fully neural object detection solutions for robot soccer." *Neural Computing and Applications* 34, no. 24 (2022): 21419-21432.

[22] Hirchoua, Badr, and Saloua El Motaki. "Semantic segmentation with transfer learning for self-driving cars." Artificial Intelligence of Things in Smart Environments: Applications in Transportation and Logistics (2022): 63.

[23] Shin, Seokyong, SangHun Lee, and HyunHo Han. "A Study on Residual U-Net for Semantic Segmentation based on Deep Learning." Journal of Digital Convergence 19, no. 6 (2021): 251-258.

[24] Ahmed, Ifham Abdul Latheef, and Mohamed Hisham Jaward. "Classifier aided training for semantic segmentation." Journal of Visual Communication and Image Representation 78 (2021): 103177.

[25] Zhang, Jing, Shaofu Lin, Lei Ding, and Lorenzo Bruzzone. "Multi-scale context aggregation for semantic segmentation of remote sensing images." Remote Sensing 12, no. 4 (2020): 701.

[26] Sushma Jaiswal, Harikumar Pallthadka, Rajesh P. Chinchewadi, & Tarun Jaiswal. (2024). A Deep Learning Model for Automatic Image Captioning using GRU and Attention Mechanism. International Journal of Computer Engineering in Research Trends, 11(1), 28–36.

[27] Sushma Jaiswal, Harikumar Pallthadka, Rajesh P. Chinchewadi, & Tarun Jaiswal. (2023). An Extensive Analysis of Image Captioning Models, Evaluation Measures, and Datasets . International Journal of Computer Engineering in Research Trends, 10(12), 31–41.

[28] Sushma Jaiswal, Harikumar Pallthadka, Rajesh P. Chinchewadi, & Tarun Jaiswal. (2023). Enhancing Image Descriptions with Image Transformers: A Journey into Advanced Image Captioning. International Journal of Computer Engineering in Research Trends, 10(12), 12–21.

[29] Venkata Srinivasu Veesam, & Bandaru Satish Babu. (2017). A Relative Study on the Segmentation Techniques of Image Processing. International Journal of Computer Engineering in Research Trends, 4(5), 155–160.