**PaveGuard: Introducing a Novel U-Net Architecture for Multi-class Segmentation in Urban Environments**

Jyoti Madake[1], Sahish Pandav[2], Om Tidke [3], Madhura Pande[4]   
*Department of Electronics and Telecommunication Engineering   
Vishwakarma Institute of Technology* Pune, India

{jyoti.madake, sahish.pandav21, baban.om21, madhura.pande211} @vit.edu

Image segmentation is a way of dividing a digital image into subgroups called image segments, reducing the complexity of the image. Segmentation is the assignment of labels to pixels to identify objects, people, or other important elements in the image. Semantic segmentation in real time is crucial for applications pertaining to mobility of robots. This paper investigates the challenging subject of street picture segmentation for self-driving cars and food delivery robots utilizing the CamVid dataset. The paper proposes use of residual neural networks within the complex UNet architecture. With a very smaller number of trainable parameters the model becomes lightweight, making it suitable for embedded systems with limited resources. The suggested solution achieves a commendable accuracy rate of 88.9%, attained through training with a minimal number of trainable parameters. Additionally, the mean Intersection over Union (mIoU) of the model is measured at 51.19. In today's world, as autonomous vehicles are rapidly becoming more popular, emphasizing the significance of effective image segmentation is crucial. The advancements presented in this paper contribute to the ongoing efforts in enhancing the capabilities of autonomous systems, ensuring their efficacy in real-world scenarios.

1. Introduction

Deep learning has been used in various fields, leading to the rapid development of image segmentation models. These models have been employed in recent years for precision agriculture, disaster management, automation using drones, automated vehicles, etc. The need for better accuracy and ever evolving technology has led to many optimizes techniques for image segmentation namely, encoder-decoder architectures, multiscale and pyramid-based approaches, recurrent networks, visual attention models, and generative models in adversarial settings form the foundation of convolutional pixel-labeling networks. An examination of the development raises the need for not only better accuracy, but for a more lightweight model [1].

Pattern recognition has led to various segmentation techniques like threshold segmentation, edge detection, etc [2]. This has led to image segmentation being useful for fire prediction and detection, biomedical, remote sensing, transportation, and engineering. The underpinning challenge lies in the embedded. The underpinning challenge lies in the embedded systems, demanding not only high accuracy in processing but also swift computation times to ensure real-time responsiveness [3]. This dual requirement underscores the necessity for models that strike an optimal balance between precision and computational efficiency in the context of embedded systems.

Modern model development mandates a departure from singularly pursuing high accuracy, necessitating a simultaneous emphasis on lightweight design. This shift in paradigm acknowledges that optimal model performance demands both precision and computational efficiency. Particularly in resource-constrained environments, models must delicately balance accuracy with a streamlined architecture to ensure operational expediency. This nuanced perspective underscores the essential interplay between model accuracy and weight in the discourse surrounding contemporary model architectures.

This paper introduces a novel UNet architecture which was motivated by the UNet framework and the power of Deep Residual Learning. The goal is to achieve Multiclass semantic segmentation in urban environments to help in faster computation of space. To achieve this goal, the project leveraged the Cambridge-driving Labeled Video Database (CamVid), which offers a unique collection of high-quality videos with object class semantic labels, complete with metadata. This dataset was chosen due to its relevance to urban scenarios and the diverse range of object classes observed from a driving automobile's perspective. With over 700 manually annotated images and detailed calibration sequences, the CamVid dataset proved invaluable in training and evaluating the model. Multi-obstacle detection in urban environments presented several challenges, including the need for high precision and accuracy.

The enhanced ResUNet architecture, an innovation built upon the UNet framework and Deep Residual Learning, played a pivotal role in addressing these challenges. ResUNet's fully convolutional neural network design enabled high performance while efficiently managing parameters, making it suitable for the demands of urban scene understanding.

The principal components of the proposed model's architectural framework encompass multi-scale feature extraction and context aggregation, both of which serve to augment the accuracy of obstacle detection. The remarkable ability of ResUNet to capture contextual information and adapt to diverse object classes represented a transformative advancement that significantly improved the segmentation results.

1. Related work

The literature has delved deeply into the issue of semantic segmentation in urban settings. Earlier methods for solving this issue focused on conventional computer vision methods like object recognition and edge detection. Deep learning has become a viable method for obstacle identification in recent years, leveraging advanced neural network architectures to achieve remarkable accuracy in segmenting urban scenes.

Pre-deep learning era, semantic segmentation approaches utilized graphical models like Markov Random Fields [8] or Conditional Random Fields [9] to capture pixel-level scene labels by modeling dependencies among neighboring pixels. However, these methods faced challenges in abstract exploitation for large-scale data. Deep learning, particularly Fully Convolutional Networks (FCNs), revolutionized semantic segmentation by adapting classification networks into fully convolutional architectures, eliminating the need for fully connected layers and significantly speeding up inference [10]. FCNs introduced skip architectures, preserving information through skip connections, enabling pixel-wise predictions for images of any resolution. Almost all subsequent semantic segmentation approaches followed the FCN paradigm, rendering fully connected layers obsolete in this context.

However, FCNs had drawbacks, including inefficient label localization, limited global context knowledge, and a lack of multiscale processing [11]. Subsequent studies focused on addressing these issues, leading to the development of various architectures and techniques. One such architecture introduced was UNet introduced in 2015[12]. The convoluted neural network-based architecture proposed a contracting path for context capture and an expansive path for precise localization. Leveraging heavy data augmentation, the network exceeded in training with minimal annotated samples, outperforming previous methods. With 23 convolutional layers, the network allows seamless tiling of the segmentation map. Overall, the approach addressed the challenge of effective deep network training with limited annotated data.

A method called ‘residual learning’ was introduced by K. He., to address challenges in training an extremely deep neural network [13]. By reformulating layers to learn residual functions with reference to layer inputs, the proposed framework makes it simpler to optimize and achieve higher accuracy with increased depth. Experiments on the ImageNet dataset and other empirical data show that residual networks with depths up to 152 layers are more effective than earlier models. ShuffleSeg, proposed by M. Gamal, et, introduced a computationally efficient real-time semantic segmentation network designed for mobile and robotics applications [14]. To improve performance, the architecture uses channel shuffling and grouped convolution in its encoder. Research comparing the decoding techniques—Skip architecture, UNet, and Dilation Frontend—shows that Skip architecture achieves the optimum compromise between accuracy and real-time performance. ShuffleSeg was tested on CityScapes.

In a discussion about deep residual networks, K. He, et al [15] discusses deep residual networks, emphasizing their efficacy in achieving high accuracy and favorable convergence behaviors. The analysis focuses on the propagation formulations within residual building blocks, showing the direct transmission of signals between blocks through identity mappings and post-addition activation. A novel residual unit that is intended to improve generalization and facilitate training is proposed because of the findings.

An approach to enhance semantic segmentation networks, focusing on the impact of convolutional layer count, is discussed by J. Liao, et al [16]. Layer count hyper-parameters are presented along with the design of loss weights for depth supervision branches that have quadratic or linear relationships with layer count. Using PSV and CamVid datasets, this technique improves medium-sized networks dramatically and increases MIoU scores by 0.94% (CamVid) and 1.37% (PSV). Significant performance improvements are shown for shallow networks; on the PSV and CamVid datasets, for example, MIoU scores rise by around 1% and 8%, respectively. Findings on Cityscapes and CamVid confirm that deep supervision enhances learning without introducing new parameters or changing the architecture of the backbone network.

The critical risk of semantic segmentation in road scenes for autonomous driving systems is addressed by Kumar R. et al [17]. The author presents a new deep learning architecture called PNet and evaluates it against cutting-edge models like FCN, SegNet, and UNet. With an 80.9% Dice Coefficient measure (DCM) and a 70.7% mean Intersection over Union (mIoU) on the CamVid dataset, the suggested PNet is trained end-to-end and outperforms current techniques. The outcomes show how well PNet performs in precisely categorizing and segmenting objects in road scenarios, which advances the field of autonomous vehicle technology. Further on, a novel strategy for semantic segmentation in complex scenes by introducing a multi-layer convolutional sparse coding block is proposed by H. Tang., et al. When this method is used with the U-Net model, it accelerates convergence, improves the extraction of semantic and appearance information, and improves the recovery of spatial detail. On the DeepCrack (87.14% vs. 84.71%), Nuclei (68.91% vs. 67.09%), and CamVid (53.68% vs. 48.82%) datasets, the top CSC-Unet model performs better than the original U-Net. [18]

The explored papers showcase significant progress in semantic segmentation and deep learning, with models like deep residual networks and convolution neural networks at the forefront. However, their computational demands pose challenges for deployment in resource-constrained or real-time mobile and robotic applications. Sensitivity to hyperparameter choices, especially regarding convolutional layer count, adds complexity for generalization. Focusing on specific challenges, such as road scene segmentation, raises concerns about overlooking broader real-world complexities..

1. Methodology

The paper presents a novel approach for feature extraction using Enhanced ResUnet. The proposed architecture is built upon the UNet framework and Deep Residual Learning. The preprocessed training images are run through the Enhanced ResUnet model to get segmented image output. The flow of the entire project is shown in Fig. 1.

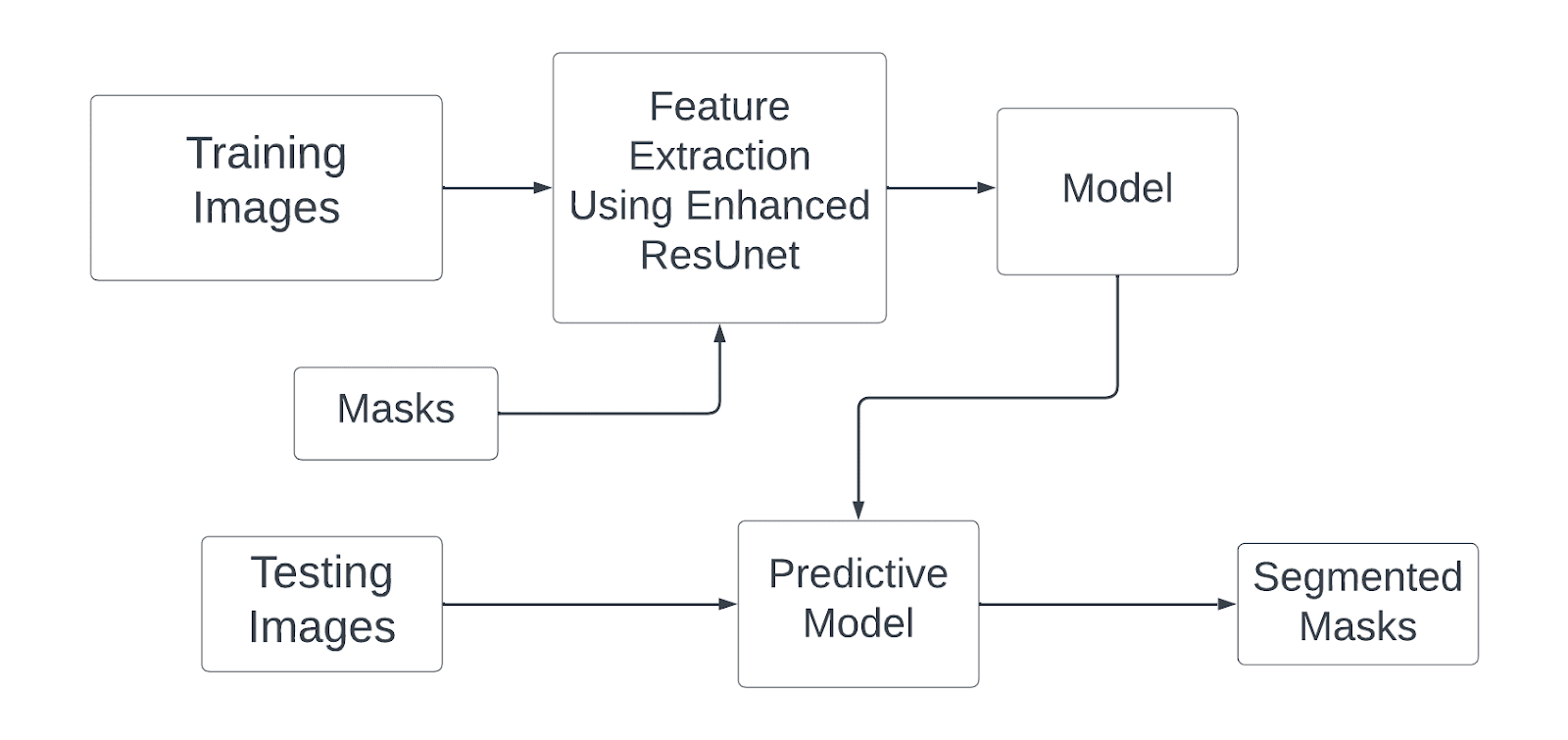


Fig. 1. Block diagram of the enhanced image segmentation

* 1. Dataset Pre-processing

The dataset employed in this study is CamVid, also known as the Cambridge Tagged Driving Video Database [5,6,7]. CamVid comprises five video sequences capturing road and driving scenarios, originally recorded using a 960 × 720 quality camera mounted on a car dashboard. As depicted in Fig. 2a, an example image from the dataset is shown, while Fig. 2b displays the corresponding mask. This database offers ground truth labels assigning each pixel to one of 32 semantic classes, including various elements such as traffic cones, bridges, signs, text, buildings, trees, and more.

|  |  |
| --- | --- |
| (a) | (b) |

Fig. 2 Sample Images from Dataset: (a) images (b) corresponding mask

The original photos in this dataset were subjected to a resizing method that entailed adding padding of 32 horizontally and 24 vertically to obtain a consistent size and allowed the image dimensions to be split equally into 256x256 patches desirable for the models input dimensions. The least occurring classes are being deleted from the dataset replacing them with Null class resulting in a total of 27 majorly occurring classes left in the prepared dataset. This technique created a total of 8000+ distinct images as well as corresponding masks from the original dataset. These produced sub-images were treated as independent entities in both training and validation stages.

* 1. Feature Extraction of Urban Environment using U-Net Architecture

To achieve detailed results in segmentation needs considering both fine grained features and large level semantic information. However training networks for such tasks might be tough, particularly with low training data. One successful way to tackle this difficulty is to employ trained networks and fine tune them on the target dataset, a regularly used technique in semantic segmentation. Another option includes applying data augmentation techniques as exemplified by the UNet architecture.

The architecture of U-Net [11, 19, 20, 21] not only benefits from data augmentation but also serves a role in aiding training. Its architecture enables for the propagation of information from low-level characteristics to comparable levels making it easier for propagation during training. Moreover, this approach helps merge low-level information with level characteristics compensating for finer details. This architectural aspect shows a resemblance to networks.

In the U-Net architecture, a simple convolutional unit performs post activated convolution, convolution — ReLu — convolution — ReLu shown in Fig. 3.

A diagram of a process

Description automatically generated

Fig. 3 Enhanced Feature Extraction using Residual Block

* 1. Enhanced Feature Extraction using Residual Block

A multilayer neural network's performance would rise with deeper layers, but doing so could make training more difficult and lead to degradation issues. To address these issues, the residual neural network handles the degradation issue and makes movement easier. A series of stacked residual units make up the residual neural network. The formula for residual networks is y = f(x) + x. where f(x) is the number of layers of the BN — ReLu — convolution, typically two of which are also regarded as pre-activated residual units in Fig. 3, and x is the input to the residual unit.

The computation carried out by the pre-activated Residual Unit is as follows:

(1)

Equation (1), the l-th Residual Unit's input feature is denoted by . Here is a Residual function shown in Fig. 4.  represents the set of weights connected to the l-th Residual Unit, and denotes the number of layers (two) in a Residual Unit. is an identity function showing feature flow from input to addition.[15]

(2)

In (2), the original input is directly added without passing through any preceding functions. The input features do not undergo any function before addition.

Therefore, substituting (2) in (1)

(3)

Recursively, the output after residual units will be

(4)

Here, shows the output of two residual blocks connected in series.

Further simplifying (4)

(5)

In (5),  XL is the output for any deeper unit L and any shallower unit l. [15]

A diagram of a computer program

Description automatically generated

Fig. 4 Residual Block

Figures 3 and 4 illustrate how a plain and residual unit differ from one another. A residual unit contains multiple combinations of convolutional layers, batch normalization (BN), and ReLU activation. By combining the initial input with the output, the addition operation makes it possible to create a residual channel for information flow. Additionally, the input is kept intact as the addition's result is passed straight into the following layer. This improves the segmentation performance by allowing the network to efficiently acquire and transfer fine-grained features from low to high levels. [16] With ultra-deep networks, this also leads to enhanced accuracy, stability, and performance.

* 1. Enhanced Feature Extraction using Residual Block

The Enhanced ResUnet Architecture for semantic segmentation, combines the best features of residual learning and U-Net architecture, gain two advantages from this combination: A neural network with significantly fewer parameters can be designed and still achieve comparable, ever-better semantic segmentation performance due to two factors: 1) The residual unit simplify network training; 2) Information propagation between low and high levels of the network and within residual units will be facilitated by skip connections without degradation.

The block diagram of ResUnet employs a seven-tier architecture for feature extraction in urban environments Fig. 4 The three components of the network are the encoder, bridge, and decoder elements. The input image is encoded into compact representations in the first section. The representations are reconstructed for semantic segmentation or pixel-wise categorization in the final stage. The center section acts as a link between the pathways for encoding and decoding. Two 3 × 3 convolution layers and an identity mapping make up the residual units used in the construction of each of the three sections. A BN layer, a ReLU activation layer, and a convolutional layer are present in every residual block. The identity mapping connects the unit's input and output. The three units make up the encoding route. Each unit applies a stride of 2 to the first convolution block to decrease the feature map by half, instead of utilizing a pooling operation to downsample the feature map size. Likewise, there are three residual units in the decoding route. Before every unit, the feature maps from the relevant encoding path are concatenated with an upsampled version of the lower-level feature maps. In the final stage of the decoding process, the multi-channel feature maps are projected onto the intended segmentation using a 1 x 1 convolution and a sigmoid activation layer. Compared to U-Net's 23 layers, the proposed model has a total of 15 convolutional layers.

A diagram of a bridge

Description automatically generated

Fig. 5 Block Diagram of Enhanced ResUnet Architecture

The initial part of the Enhanced ResUNet is an encoder, whose function it is to extract features from the input image as shown in Fig. 5. Each level applies multiple filter layers to the original image in order to capture increasingly abstract qualities. The spatial dimensions of the feature maps decrease as their depth increases. Repetition of this method yields more feature channels and less spatial resolution. They facilitate the network's focus on more important patterns. A residual unit, each of the three encoder blocks (level 1 through level 3) is made up of two convolutional layers and two activation layers. This improves the capturing of the hierarchical features of the supplied image.

Following the encoder comes a bottleneck layer (level 4), which records very high-level characteristics. As a bridge, it connects the encoder and decoder. The decoder upsamples the feature maps to the resolution of the original input picture while progressively refining the segmentation mask. To prevent the feature map's size from decreasing, each of the three decoder blocks (levels 5 through 7) begins with transposed convolutional layers, often referred to as fractionally strided convolution or deconvolution, and is followed by a residual unit with (stride = 1). These layers carry out upsampling by learning to reverse the impact of pooling operations. Fig. 6 shows the detailed architecture of the proposed model highlighting residual units and skip connections between encoder and decoder.

The upsampled feature maps from the transposed convolutional layers are concatenated with the matching feature maps from the encoder in the decoder, which uses skip connections. This is what the concatenate layer does. Skip links aid in maintaining features and geographic information that are lost during the encoder's downsampling procedure. They enable the network to efficiently blend high-level and low-level features.

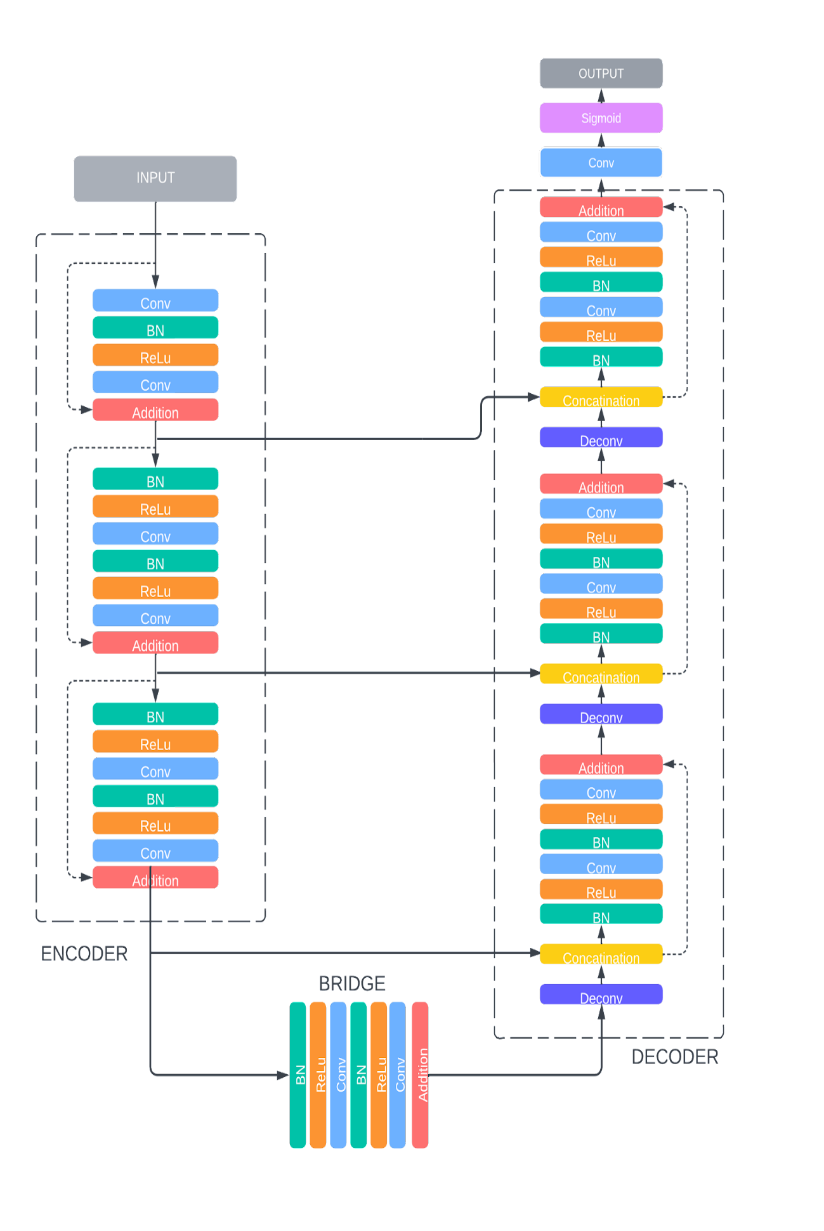


Fig. 6  Detailed Diagram of Proposed Architecture

A multiclass segmentation mask is produced by a crucial convolutional layer with 27 filters and a softmax activation algorithm. Softmax guarantees a normalized probability distribution for fine-grained pixel-by-pixel classification, whereas these filters capture a variety of features. The resulting mask provides in-depth information on the image's content, with each pixel corresponding to a class label. This customized architecture demonstrates its versatility in identifying and classifying complex visual features for accurate scene analysis in computer vision, as demonstrated by its success in applications such as semantic and instance segmentation.

* 1. Evaluation Metrics

Throughout the iterative training process, the dynamics of model convergence and generalization performance are gauged through the scrutiny of training and validation loss curves. These curves serve as indispensable tools for monitoring the model's learning dynamics and identifying potential challenges, such as underfitting or overfitting [23, 24]. The training loss curve, mapping the progression of categorical cross-entropy loss over training epochs, is expected to exhibit a diminishing trend as the model converges to an optimal fit for the training set. Notably, the dataset was partitioned into an 80%-20% split for training and testing, ensuring a robust evaluation of the model's performance on previously unseen data.

Similarly, the validation loss curve portrays the model's performance on an independent validation dataset, offering insights into its capacity to generalize to new, unseen data. Effective learning without overfitting is discerned through a diminishing or stabilizing validation loss trend. The evaluation framework extends to the mean Intersection over Union (mIoU) metric, as denoted by formula (6). This metric quantifies the model's ability to accurately segment and classify objects across diverse classes. The formula, averaging the IoU values for each class, is expressed as:

(6)

In this context, N represents the total number of classes, and TPi​, FPi​, and FNi​ stand for the counts of correctly identified (true positive), wrongly identified as positive (false positive), and wrongly identified as negative (false negative) pixels for a particular class i. To improve how well the model works, we use techniques like early stopping, adjusting the learning rate, and tuning hyperparameters. Important metrics, such as categorical cross-entropy and mIoU, help us fine-tune the model for its best performance during both training and validation. Regularly checking these curves helps us figure out if the model has learned well enough or if it needs further adjustments through more training iterations [25-29].

1. Results and Discussion

The Enhanced ResUnet architecture demonstrated superior performance when compared to both the original UNet and other segmentation methods applied to the dataset. Evaluation metrics, including the categorical cross entropy, and mean Intersection over Union (mIOU), were employed to assess the segmentation performance.

A graph of training and validation loss

Description automatically generatedA graph of a training and validation accuracy

Description automatically generated  
 (a) (b)

Fig. 7 (a) Training and Validation loss for CamVid dataset over 100 epochs. (b) Training and validation accuracy for CamVid dataset over 100 epochs.

In Fig. 7a there is a consistent decrease in both training and validation losses over the course of 100 epochs, this is generally a positive sign during the training of a segmentation model.

Table 1: Training Results

|  |  |
| --- | --- |
| Evaluation Matrix | (100 epoch) |
| Training Loss | 0.0634 |
| Training Accuracy | 0.9776 |
| Validation Loss | 0.2251 |
| Validation Accuracy | **0.8888** |
| Mean IoU | 0.5119 |

Table 1 shows the results of training Enhanced ResUnet on CamVid dataset. After 100 epochs, training ceased due to stagnant accuracy metrics. The 80%-20% dataset split ensured robust evaluation. The model likely reached a global minima, representing the lowest possible loss in the optimization process.

The model's training accuracy of almost 98% (Table 1) was one of the encouraging results of the training phase. This high accuracy shows that the complex patterns and characteristics found in the training dataset were successfully mastered by the ResUNet architecture. Concurrently, the corresponding training loss shows that the model was successfully optimized during the training phase, converging to approximately 6%. As the network processes the training data iteratively, the decreasing training loss is a sign that the network is improving its predicting accuracy and minimizing errors.

A screenshot of a computer generated image

Description automatically generated

A screenshot of a video game

Description automatically generated

A screenshot of a computer generated image

Description automatically generated

Fig. 8 Semantic Segmentation Results

This model demonstrated an impressive 88% validation accuracy (Fig. 7b). This demonstrates the model's proficiency in classifying diverse objects or features in unknown data. Caution is necessary for potential issues affecting segmentation quality, like unequal class representation or complex real-world circumstances. These criteria are crucial in evaluating the model's performance.

The visual similarity between the model's predicted masks and the reference masks in Figure 8 provides evidence for the segmentation model's accuracy and reliability. The model's abilities to build masks that virtually match the actual data illustrates how well it can capture fine-grained details and semantic boundaries. The fact that the predicted and original masks align emphasizes the model's capacity to reliably replicate the semantic content of the input images and verifies the quantitative metrics, such as the mean Intersection over Union (mIoU) score of 51.19.

Table II: Comparison with existing segmentation models

|  |  |  |  |
| --- | --- | --- | --- |
| Model | mIoU | Parameters | Model Size (MB) |
| U-Net[6] | 36.59 | 34,538,912 | 48.32 |
| SegNet[29] | 45.89 | 71,744,025 | 117 |
| PSPNet(269)[6] | 43.81 | 53,438,912 | 162 |
| Proposed Model | 51.19 | 5,485,002 | 50.92 |

When compared to other well-known models (Table 2) in the field of semantic segmentation, Enhanced ResUnet model not only performs competitively but also stands out for its efficacy and streamlined architecture. With an mIoU of 51.19, the model outperforms UNet architecture with a greater mIoU, and performs on par with SegNet[29].

Most critically, the paradigm differs in that it prioritizes efficiency. When compared to its competitors, Enhanced ResUnet has a minimal number of trainable parameters. It is nevertheless effective in spite of its efficiency, illustrating the harmony between precision and the use of computer resources. With a modest architecture consisting of only 15 convolution layers, it illustrates that enhanced segmentation results can be produced without excessive complexity increase. This contrast underlines the model's applicability and makes it especially pertinent for applications where computational resources are a significant concern. In the future, more excellent optimization and fine-tuning could make the model even more competitive in assessments to come.

1. Conclusion

PaveGaurd exhibits broad potential use in embedded systems that require low resources. It optimized the U-Net architecture by using residual networks, which enhanced the attention mechanisms. Deep residual networks excel when information flows seamlessly through the network layers, facilitating better gradient propagation and easing training. Similarly, Unet demonstrates superiority in semantic segmentation by effectively acquiring hierarchical features. When these structures collaborate, an efficient model is formed, showcasing versatility, interpretability, and efficacy in picture segmentation applications. The ResUnet, combining deep residual networks with Unet, demonstrates its potential as a useful tool. To enhance performance, the intention is to fine-tune the model by adjusting settings, aiming to improve image segmentation accuracy. The goal is to better adapt the model to diverse scenarios, enhancing overall functionality and increasing adaptability in real-world applications.

REFERENCES

1. S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz and D. Terzopoulos, "Image Segmentation Using Deep Learning: A Survey," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 44, no. 7, pp. 3523-3542, 1 July 2022, doi: 10.1109/TPAMI.2021.3059968.
2. Cheng, Y., & Li, B. (2021). “Image Segmentation Technology and Its Application in Digital Image Processing”, 2021 IEEE Asia-Pacific Conference on Image Processing, Electronics and Computers (IPEC). doi:10.1109/ipec51340.2021.9421206
3. Y. Zhang, R. Yang, J. Wang, N. Chen and Q. Dai, "The Impact of Parameters on Semantic Segmentation: A Case Study on the CamVid Dataset," 2021 IEEE 23rd Int Conf on High Performance Computing & Communications; 7th Int Conf on Data Science & Systems; 19th Int Conf on Smart City; 7th Int Conf on Dependability in Sensor, Cloud & Big Data Systems & Application (HPCC/DSS/SmartCity/DependSys), Haikou, Hainan, China, 2021, pp. 1932-1938, doi: 10.1109/HPCC-DSS-SmartCity-DependSys53884.2021.00289.
4. Z. Zhang, Q. Liu and Y. Wang, "Road Extraction by Deep Residual U-Net," in IEEE Geoscience and Remote Sensing Letters, vol. 15, no. 5, pp. 749-753, May 2018, doi: 10.1109/LGRS.2018.2802944.
5. Cambridge University Engineering Department. "CamVid - Video Datasets for Urban Scene Understanding." https://mi.eng.cam.ac.uk/research/projects/VideoRec/CamVid/ (Accessed : 2nd November 2023)
6. Brostow, Gabriel J., Jamie Shotton, Julien Fauqueur, and Roberto Cipolla. "Segmentation and Recognition Using Structure from Motion Point Clouds." In ECCV (1), 44-57, 2008
7. Brostow, Gabriel J., Julien Fauqueur, and Roberto Cipolla. "Semantic Object Classes in Video: A High-Definition Ground Truth Database." Pattern Recognition Letters xx, no. x (2008): xx-xx
8. Y. Boykov, O. Veksler and R. Zabih, "Markov random fields with efficient approximations," Proceedings. 1998 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No.98CB36231), Santa Barbara, CA, USA, 1998, pp. 648-655, doi: 10.1109/CVPR.1998.698673.
9. D. Y. Liliana and C. Basaruddin, "A review on conditional random fields as a sequential classifier in machine learning," 2017 International Conference on Electrical Engineering and Computer Science (ICECOS), Palembang, Indonesia, 2017, pp. 143-148, doi: 10.1109/ICECOS.2017.8167121.
10. J. Long, E. Shelhamer and T. Darrell, "Fully convolutional networks for semantic segmentation," 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 2015, pp. 3431-3440, doi: 10.1109/CVPR.2015.7298965.
11. Ulku, Irem, and Erdem Akagündüz. "A Survey on Deep Learning-based Architectures for Semantic Segmentation on 2D Images." Applied Artificial Intelligence 36, no. 1 (2022): doi: 10.1080/08839514.2022.2032924.
12. Ronneberger, O., Fischer, P., & Brox, T. (2015). "U-Net: Convolutional Networks for Biomedical Image Segmentation." In N. Navab, J. Hornegger, W. Wells, A. Frangi (Eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. (Vol. 9351, Lecture Notes in Computer Science). Springer, Cham. https://doi.org/10.1007/978-3-319-24574-4\_28
13. K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 770-778, doi: 10.1109/CVPR.2016.90.
14. M. Gamal, M. Siam, and M. Abdel-Razek. "SHUFFLESEG: REAL-TIME SEMANTIC SEGMENTATION NETWORK," March 15, 2018. https://doi.org/10.48550/arXiv.1803.03816.
15. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Identity Mappings in Deep Residual Networks," July 26, 2016. https://doi.org/10.48550/arXiv.1603.05027.
16. J. Liao, L. Cao, W. Li, Y. Ou, C. Duan and H. Cao, "Fully-Supervised Semantic Segmentation Networks: Exploring the Relationship between the Segmentation Networks Learning Ability and the Number of Convolutional Layers," 2021 7th International Conference on Computer and Communications (ICCC), Chengdu, China, 2021, pp. 1685-1692, doi: 10.1109/ICCC54389.2021.9674634.
17. Kumar, R., Amitab, K. (2023). “Semantic Segmentation of Road Scene Using Deep Learning. “In: Basu, S., Kole, D.K., Maji, A.K., Plewczynski, D., Bhattacharjee, D. (eds) Proceedings of International Conference on Frontiers in Computing and Systems. Lecture Notes in Networks and Systems, vol 404. Springer, Singapore. <https://doi.org/10.1007/978-981-19-0105-8_13>
18. H. Tang, S. He, X. Lu, Q. Yu, K. Liu, H. Yan and N. Wang, "CSC-Unet: A Novel Convolutional Sparse Coding Strategy based Neural Network for Semantic Segmentation", <https://doi.org/10.48550/arXiv.2108.00408>
19. Siddique, N., Paheding, S., Elkin, C. P., & Devabhaktuni, V. (2021). “U-Net and Its Variants for Medical Image Segmentation: A Review of Theory and Applications” IEEE Access, 9, 82031–82057. doi:10.1109/access.2021.3086020
20. [20]Deng, Yunjiao & Hou, Yulei & Yan, Jiangtao & Zeng, Daxing. (2022). “ELU-Net: An Efficient and Lightweight U-Net for Medical Image Segmentation. IEEE Access. 10. 1-1. 10.1109/ACCESS.2022.3163711.
21. Wei, Song & Zheng, Nan & Liu, Xiangchun & Qiu, Lirong & Zheng, Rui. (2019). An Improved U-Net Convolutional Networks for Seabed Mineral Image Segmentation. IEEE Access. 7. 1-1. 10.1109/ACCESS.2019.2923753.
22. Anand, Vatsala & Gupta, Sheifali & Ko, Deep & Nayak, Soumya & Barsocchi, Paolo & Bhoi, Akash Kumar. (2022). Modified U-NET Architecture for Segmentation of Skin Lesion. Sensors. 22. 867. 10.3390/s22030867.
23. Road segmentation using U-NET design from satellite imagery Ghosh, Souvik Said, Yahia, Mohamed Atri, Marwan Ali Albahar, Ahmed Ben Atitallah, and Yazan Ahmad Alsariera. 2023. "Obstacle Detection System for Navigation Assistance of Visually Impaired People Based on Deep Learning Techniques" Sensors 23, no. 11: 5262. https://doi.org/10.3390/s23115262
24. Hussain, S. S., Durrani, D., Khan, A. A., Atta, R., & Ahmed, L. (2020). In-door Obstacle Detection and Avoidance System for Visually Impaired People. 2020 IEEE Global Humanitarian Technology Conference (GHTC). doi:10.1109/ghtc46280.2020.9342942
25. Mane, Shraddha and Prof.Supriya Mangale. “Moving Object Detection and Tracking Using Convolutional Neural Networks.” 2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS) (2018): 1809-1813.
26. Wang, Fang & Xie, Jindong. (2020). A context and semantic enhanced UNet for semantic segmentation of high-resolution aerial imagery. Journal of Physics: Conference Series. 1607. 012083. 10.1088/1742-6596/1607/1/012083.
27. Baheti, Bhakti & Innani, Shubham & Gajre, Suhas & Talbar, Sanjay. (2020). Eff-UNet: A Novel Architecture for Semantic Segmentation in Unstructured Environment. 1473-1481. 10.1109/CVPRW50498.2020.00187.
28. M. Ehrig and J. Euzenat, “Relaxed precision and recall for ontology matching,” in Workshop on Integrating ontology, 2005, pp. 25–32.
29. V. Badrinarayanan, A. Kendall and R. Cipolla, "SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 12, pp. 2481-2495, 1 Dec. 2017, doi: 10.1109/TPAMI.2016.2644615.