Multi-UNet++: A Hybrid Model for Image Segmentation

Shreya Ghansham Bawaskar (22CSM2R19)
M.Tech 1st year CSIS
National Institute of Technology, Warangal
Telangana, India 506004

***Abstract*— Image segmentation is the process of dividing an image into a number of useful regions or segments according to specific standards or characteristics. Grouping pixels or regions with comparable attributes, such as colour, texture, intensity, or semantic meaning, is the aim. Object recognition, scene understanding, image editing, and medical imaging analysis are just a few of the applications that can be made possible by image segmentation, which is essential in computer vision tasks. In the proposed model, the hybrid model is designed from two different models such as multi-CNN model and UNet++. In the given architecture two UNet++ models with different pretrained encoders are used to learn feature sets with same class distribution. So, it becomes easy for the network to correct each other by exchanging their loss information and produces better results.**

I. INTRODUCTION

Image segmentation is the process of dividing an image into a number of useful regions or segments according to specific standards or characteristics. Grouping pixels or regions with comparable attributes, such as color, texture, intensity, or semantic meaning, is the aim. Object recognition, scene understanding, image editing, and medical imaging analysis are just a few of the applications that can be made possible by image segmentation, which is essential in computer vision tasks. A computer vision task called semantic segmentation entails giving each pixel in an image a label or category. Dense prediction, or labelling every pixel in the image with the class to which it belongs, is what is intended. Other methods of image classification or object detection, which put more of an emphasis on labelling entire objects or regions of interest than specific pixels, are different from this. Numerous applications, including scene understanding, autonomous driving, image editing, and medical image analysis, benefit from semantic segmentation. Semantic segmentation and instance segmentation are the two primary tasks in the area of image segmentation. Semantic segmentation involves assigning a class or category to each pixel in an image. The objective is to semantically classify pixels into meaningful segments. For instance, semantic segmentation would give labels to pixels in a street scene that represented the road, sky, buildings, pedestrians, vehicles, etc. The result is a dense pixel-by-pixel prediction, with a class label assigned to each pixel. With specialized architectures and research efforts, earlier segmentation approaches concentrated on tackling semantic segmentation as a stand-alone task. Instance segmentation advances the task by differentiating between specific instances or objects within the same class in addition to classifying pixels with semantic labels. It aims to recognize and categorize each instance separately, providing each object with both clear boundaries and class labels. In situations requiring accurate object localization and distinction, like object counting, tracking, or interactive image editing, this is especially helpful. Instance and semantic segmentation were typically handled separately in earlier segmentation approaches, and each was given a unique architecture. But more recent developments, such as the development of deep learning and convolutional neural networks, have resulted in the creation of integrated frameworks that can deal with both tasks concurrently and are frequently referred to as panoptic segmentation. Transformer-based vision networks are used for a variety of computer vision tasks, including image classification and semantic segmentation. These networks were inspired by the success of transformer models in natural language processing tasks. Transformer-based models have become serious competitors to Convolutional Neural Networks (CNNs), which have long been the preferred architecture for computer vision tasks. Encoder-decoder networks are frequently used in contemporary semantic and instance segmentation models. Their skip connections, which combine shallow, low-level, fine-grained feature maps from the encoder sub-network with shallow, semantic, coarse-grained feature maps from the decoder sub-network, are primarily responsible for their ability to recover fine-grained details of the target objects even against complex background. Skip connections have also been largely credited with the success of instance-level segmentation models, such as those whose objective is to segment and distinguish each instance of desired objects. In the proposed model, the hybrid model is designed from two different models such as multi-CNN model and UNet++. In the given architecture two UNet++ models with different pretrained encoders are used to learn feature sets with same class distribution. So, it becomes easy for the network to correct each other by exchanging their loss information. UNet++ is composed of U-Nets with various depths, each of whose decoders is densely connected to every other decoder at the same resolution using newly established skip connections. The high-level architecture is shown in the Fig. 1.

II. RELATED WORK

One of the most fundamental operations in image processing and computer vision is image segmentation. It is a necessary step before high resolution images can be pro-

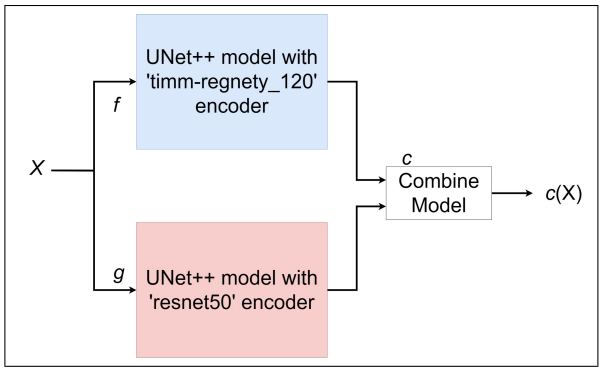

Fig. 1: Architecture of the proposed combined UNet++ Model.

cessed and analysed. For many applications, effective image segmentation is necessary when segmenting complex image scenes with obvious texture. In many image segmentation tasks across many domains, including satellite imagery, the UNet [1] network architecture has been utilized.

Convolutional neural networks (CNNs) have traditionally been used to approach semantic segmentation as a pixel classification problem. CNNs were used in early semantic segmentation studies to categorize each pixel into various semantic groups. These techniques showed good performance and made significant progress [2], [3], [4], [11].

However, more recent improvements in transformer-based techniques, motivated by transformers' success in language and vision tasks [5], [13], have also demonstrated encouraging outcomes in semantic segmentation. Transformers have been used in semantic segmentation, and researchers have achieved cutting-edge performance [10], [12], [14].

MaskFormer [7], a noteworthy method, treats semantic segmentation as a mask classification issue. The previous works [6], [8], [9] that treated segmentation as a classification problem for individual pixel masks were the foundation for this method. MaskFormer creates pixel-wise masks for various semantic categories using a transformer decoder and object queries. This formulation enables segmentation results that are more precise and thorough.

Researchers have been able to increase the efficiency and precision of semantic segmentation tasks by modeling them as mask classification problems and utilizing transformer-based architectures. These methods have shown that transformers are capable of handling spatial data and capturing long-distance dependencies, producing segmentation results that are more accurate and reliable.

III. Proposed Model

In the proposed model, the hybrid model is designed from two different models such as multi-CNN model and UNet++. In the given architecture two UNet++ models with different pretrained encoders are used to learn feature sets with same class distribution. So, it becomes easy for the network to correct each other by exchanging their loss information. UNet++ is composed of U-Nets with various depths, each

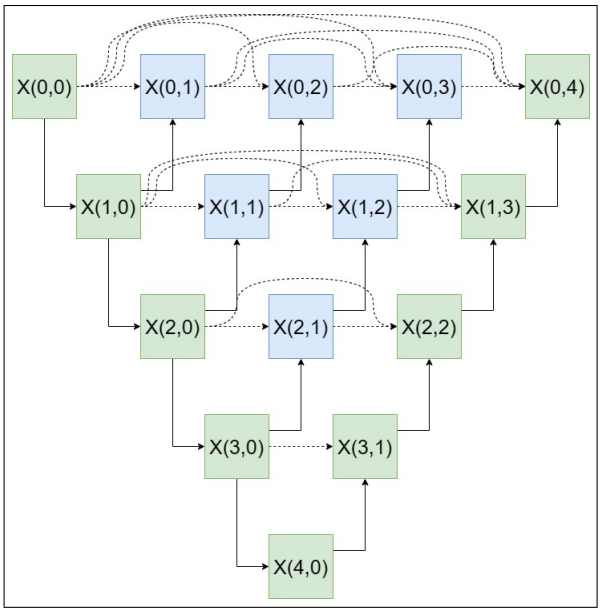

Fig. 2: UNet++ Model.

of whose decoders is densely connected to every other decoder at the same resolution using newly established skip connections shown in Fig. 2. The detailed architecture is shown in the Fig. 3. Each of the feature map generated by encoders is given to the feature fusion module. The architecture of each of the block is same which is given in the Fig. 4.

Prior to fusion, UNet++ fills in the semantic gap between the feature maps of the encoder and decoder. In UNet++, feature maps of encoder undergo dense convolutional block. The skip pathway is made up of three convolution layers in a dense convolution block, each of which is followed by a concatenation layer that fuses the output from the previous convolution layer with the corresponding up-sampled output from the lower dense block. Two UNet++ models are combined to get the better results for image segmentation problem. The combined output of encoders of two model makes it easy to classify the pixel in the correct class. In the proposed model, first UNet++ model uses 'timm-regnety_120' encoder and 'resnet50' is used in second UNet++ model. The input is passed to the encoder of both of these models. Output of these encoders will be passed to feature fusion module which combines these outputs into one by using convolutional layer of kernel size 1 and then through batch normalization layer and then Relu function is used. The output of this feature fusion model is then combined and passed to decoder and then through segmentation head. Activation function used in both model is softmax function and loss function used is DiceLoss.

IV. Implementation

A. Dataset

The proposed model is used on the DeepGlobe Land Cover Classification Dataset and Cityscapes. The DeepGlobe

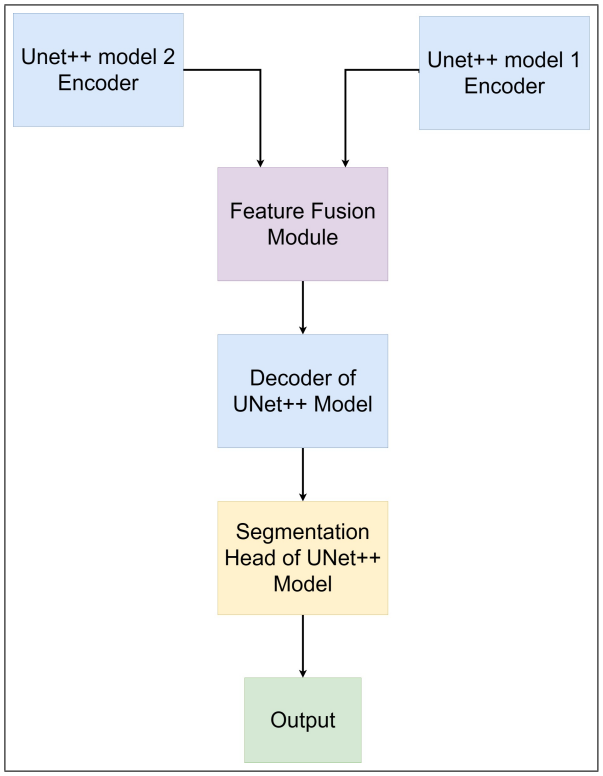

Fig. 3: Detailed Architecture of Proposed Model.

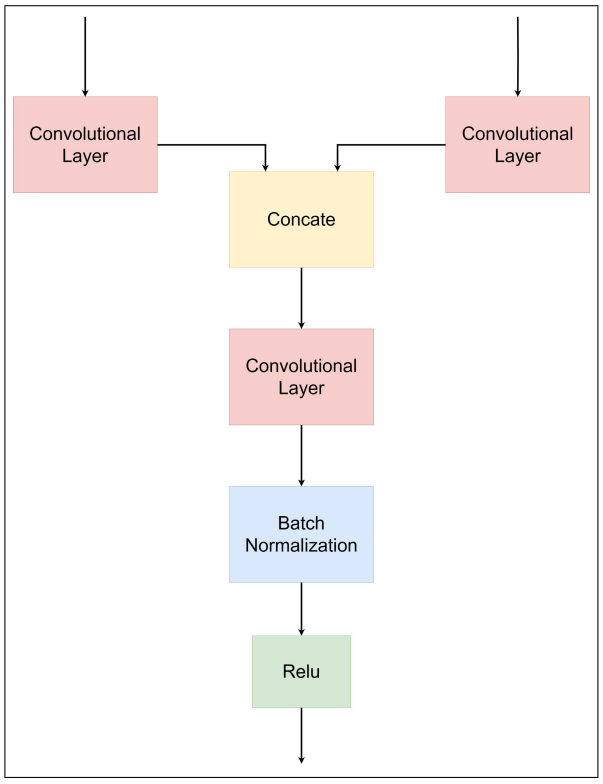

Fig. 4: Block in Feature Fusion Module.

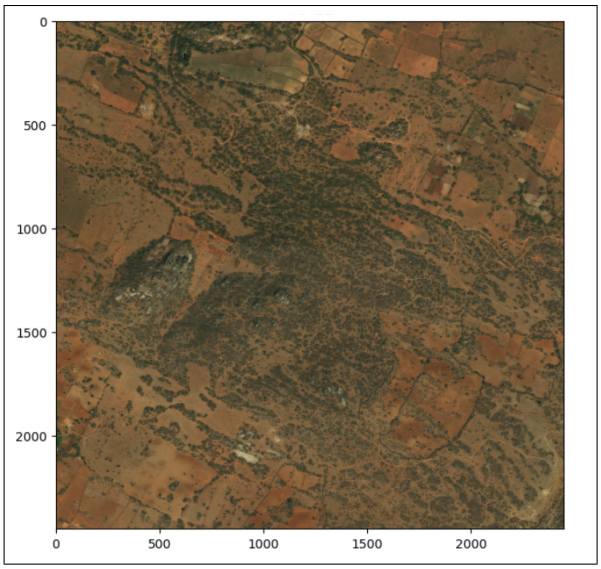

Fig. 5: Image in the dataset.

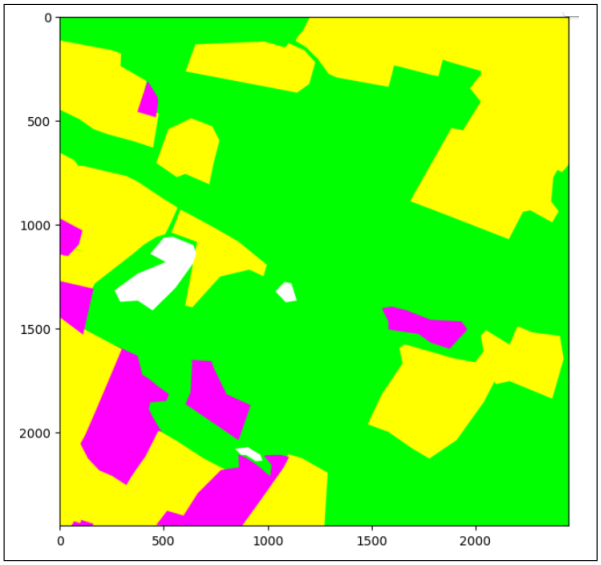

Fig. 6: Mask of image in Fig. 5.

Land Cover Classification dataset was obtained from Land Cover Classification Track in DeepGlobe Challenge. It contains RGB colour code mappings for different classes in labels. The labels in this dataset have 7 classes: 'urban-land', 'agriculture-land', 'rangeland', 'forest-land', 'water', 'barren-land' & 'unknown'. The data for Land Cover Challenge contains 803 satellite imagery in RGB, size 2448x2448. The imagery has 50cm pixel resolution, collected by DigitalGlobe's satellite. Each satellite image is paired with a mask image for land cover annotation. The mask is a RGB image with 7 classes of labels, using color-coding (R, G, B) . The image and its mask are shown the Fig. 5 and Fig. 6 respectively.

Cityscapes is a large-scale database which focuses on semantic understanding of urban street scenes. It provides semantic, instance-wise, and dense pixel annotations for 30

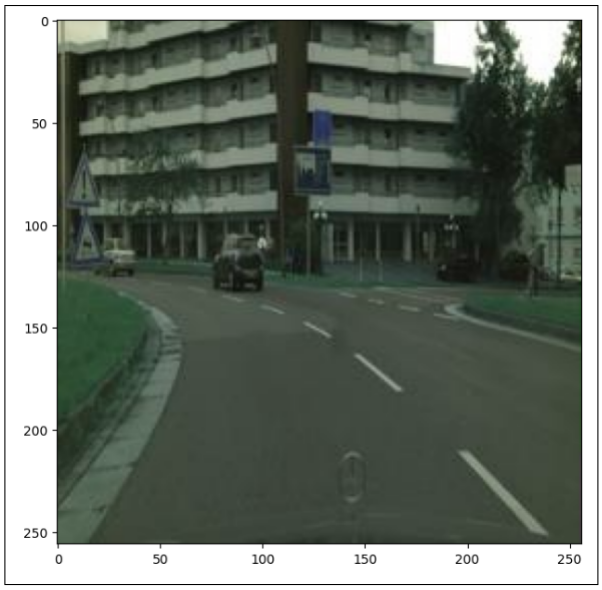

Fig. 7: Image in the dataset.

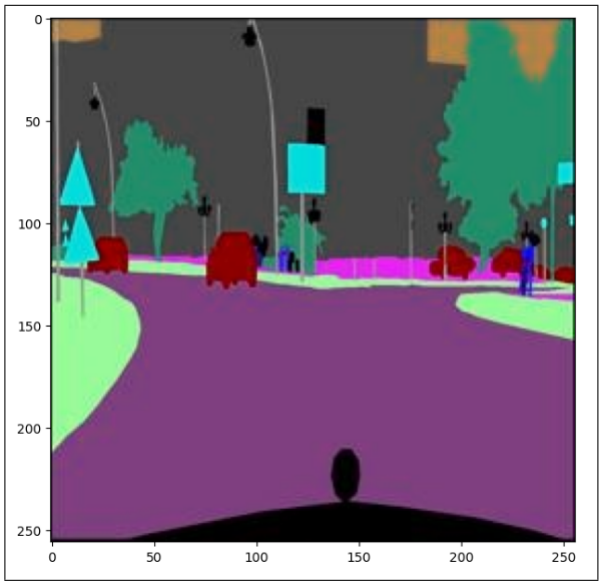

Fig. 8: Mask of image in Fig. 7.

classes grouped into 8 categories (flat surfaces, humans, vehicles, constructions, objects, nature, sky, and void). The image and its mask are shown the Fig. 7 and Fig. 8 respectively.

B. Feature Fusion Module

In this module, the feature maps produced by the encoder of each of the Unet++ model is are first passed through a convolutional layer and then the outputs are concatenated. The concatenated output is then passed through a convolutional layer and then batch normalization is applied. The output after batch normalization is given to the Relu function and then it is given as an input to the decoder of the UNet++ model and then to the segmentation head of the UNet++ model to produce the mask of the input image as shown in Fig. 4.

C. UNet Model

UNet is a fully convolutional neural network architecture that consists of a decoder network that up-samples the feature map to the original image size and an encoder network that down-samples the input image to a low-resolution feature map. Skip connections link the encoder and decoder networks, enabling the decoder to access high-resolution feature maps from the encoder at various scales.

D. UNet++ Model

By increasing the network's representation power, UNet++, an addition to the UNet architecture, seeks to improve performance. UNet++ uses a nested, or recursive, architecture, meaning that each stage in the encoder and decoder networks is made up of multiple parallel pathways as shown in Fig. 2, each of which captures a different scale of features. The skip pathway is formalized as follows: Let $x^{i,j}$ represent the output of node $X^{i,j}$ where i denotes the convolution layer of the dense block along the skip pathway and j denotes the downsampling layer along the encoder. The stack of feature maps represented by $x^{i,j}$ which is computed as follows

$$x^{i,j} = \begin{cases} \mathcal{H}(x^{i-1,j}), & \text{if } j = 0 \\ \mathcal{H}(\left[\left[x^{i,k}\right]_{k=0}^{j-1}, \mathcal{U}(x^{i+1,j-1}\right]), & \text{if} j > 0 \end{cases} \quad (1)$$

This is the main distinction between UNet and UNet++. When compared to UNet, UNet++'s segmentation results are improved by its ability to capture more varied and multi-scale features due to its nested structure. In the proposed model, one of the UNet++ model uses 'timm-regnety_120' and other uses 'resnet50' encoder. The loss function for the model used if 'DiceLoss' function and optimizer is 'AdamW'. A specific model implemented in the PyTorch Image Models (timm) library is referred to as the "timm regnety 120." It belongs to the family of RegNetY models, scalable and effective architectures made for computer vision tasks. The RegNetY-120GF model, where "GF" stands for "Growth Factor," is referred to as the "timm regnety 120" model. The modular structure of RegNet models makes it simple to scale their depth, width, and resolution. ResNet-50 uses residual blocks and a residual architecture. Shortcut connections are used by residual blocks in ResNet to address the vanishing gradient issue. Blocks called bottleneck residuals are used to lower computation costs while preserving or enhancing performance. Three layers make up a bottleneck block: one 1x1 convolutional layer, one 3x3 convolutional layer, and one more 1x1 convolutional layer. The number of channels is reduced and then restored using the first and last convolutional layers, respectively. Space-related information is captured by the middle 3x3 convolutional layer. A stochastic gradient descent method called AdamW optimization is based on adaptive estimation of first- and second-order moments with an additional method to decay weights in accordance with the techniques. It fixes the weight decay issue, works well for overfitting, and boosts generalization performance. Dice coefficient outperforms in class-imbalanced situations.

E. Training and Testing Result

All of the images and its masks are scaled to the size of 320 x 320. Horizontal and Vertical flips are applied on all of the images in training data. Both the base model and Hybrid model is trained on the same dataset with batch size of 16. Fig. 9, Fig. 10, Fig. 11 and Fig. 12 shows the evaluation on the training and validation data for both the data sets. Fig. 9 and Fig. 11 shows the IOU for proposed model on DeepGlobe Land Cover Dataset and Cityscapes dataset respectively. Fig. 10 and Fig. 12 shows the Loss for proposed model on DeepGlobe Land Cover Dataset and Cityscapes dataset respectively. Fig. **??** and Fig. 13 shows the testing results of both hybrid and base models respectively on DeepGlobe Land Cover dataset. Different models are used for the semantic segmentation on the Cityscapes dataset such as Dilated ResNet[15], PSPNet[16], Auto-DeepLab-L[17], HRNetV2[18] etc. The model which results in the highest IOU till now is InternImage-H[19] with 87% IOU. The IOU scores of these models are mentioned in Table.I The proposed model results in 99.18% IOU for Cityscapes Dataset as we can see in the Fig. 14.

Model Name	IOU Scores
Dilated ResNet	75.7%
PSPNet	79.7%
Auto-DeepLab-L	80.33
HRNetV2	81.10%
InternImage-H	87%
Proposed Model	99.18%

TABLE I: IOU scores of different models on Cityscapes Dataset.

V. CONCLUSIONS

In this report the novel architecture of combined UNet++ is presented which gives more accurate image segmentation. The improved performance of the proposed model is attributed to feature fusion of the two UNet++ models. The model is evaluated on the DeepGlobe Land Cover Classification dataset and Cityscapes dataset and demonstrated performance improvement for image segmentation.

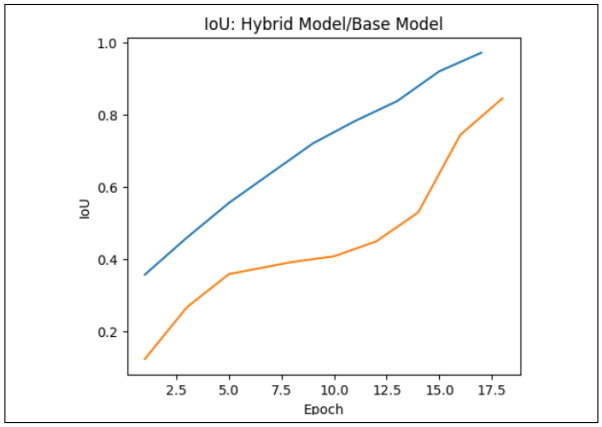

Fig. 9: IOU on DeepGlobe Land Cover Dataset

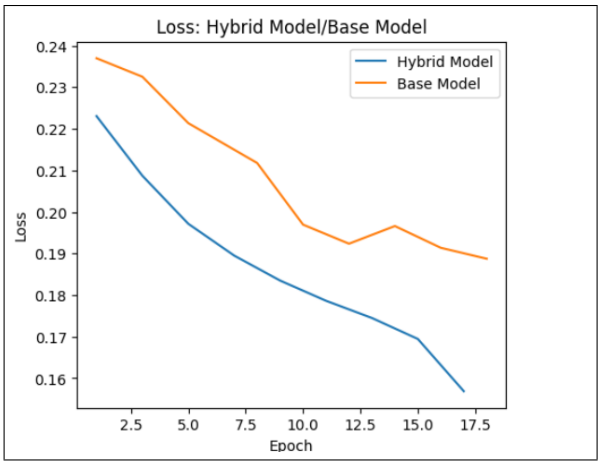

Fig. 10: Loss on DeepGlobe Land Cover Dataset

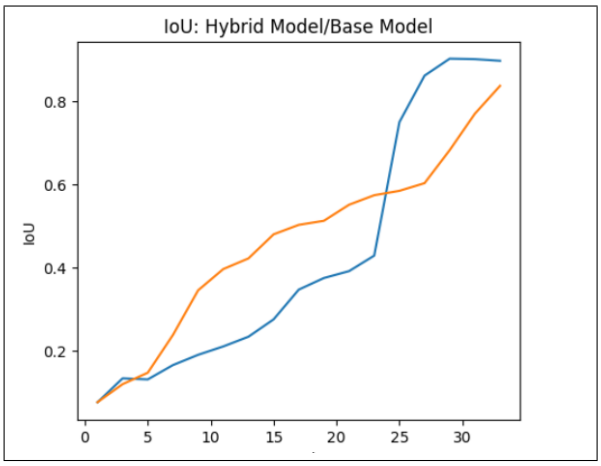

Fig. 11: IOU on Cityscrape Dataset

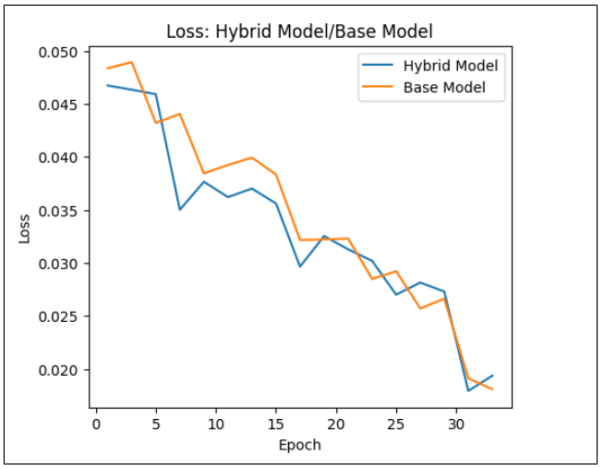

Fig. 12: Loss on Cityscrape Dataset

Test metric	DataLoader 0
test/Accuracy	0.9969748258590698
test/F1score	0.989412248134613
test/IoU	0.9806896448135376
test/Loss	0.14688684046268463
test/Precision	0.989412248134613
test/Recall	0.989412248134613

Fig. 13: Testing Result of Hybrid Model on DeepGlobe Land Cover Dataset.

Test metric	DataLoader 0
test/Accuracy	0.8807899355888367
test/F1score	0.5827645063400269
test/IoU	0.47735682129859924
test/Loss	0.19072888791561127
test/Precision	0.5827645063400269
test/Recall	0.5827645063400269

Fig. 14: Testing Result of Base Model on DeepGlobe Land Cover Dataset.

Test metric	DataLoader 0
test/F1score	0.9959042072296143
test/IoU	0.9918432235717773
test/Loss	0.012563644908368587
test/Precision	0.9959042072296143
test/Recall	0.9959042072296143

Fig. 15: Testing Result of Hybrid Model on Cityscapes Dataset.

ACKNOWLEDGMENT

I would like to thanks , Dr. P. Radha Krishna Sir and Dr. M. Srinivas Sir, Computer Science And Engineering, National Institute of Technology, Warangal, for their constant encouragement towards the realization of this work.

REFERENCES

[1] Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation." Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. Springer International Publishing, 2015.

[2] Chen, Liang-Chieh, et al. "Semantic image segmentation with deep convolutional nets and fully connected crfs." arXiv preprint arXiv:1412.7062 (2014).

[3] Chen, Liang-Chieh, et al. "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs." IEEE transactions on pattern analysis and machine intelligence 40.4 (2017): 834-848.

[4] Cheng, Bowen, et al. "Spgnet: Semantic prediction guidance for scene parsing." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019.

[5] Carion, Nicolas, et al. "End-to-end object detection with transformers." Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16. Springer International Publishing, 2020.

[6] Carreira, Joao, et al. "Semantic segmentation with second-order pooling." Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part VII 12. Springer Berlin Heidelberg, 2012.

[7] Cheng, Bowen, Alex Schwing, and Alexander Kirillov. "Per-pixel classification is not all you need for semantic segmentation." Advances in Neural Information Processing Systems 34 (2021): 17864-17875.

[8] Dai, Jifeng, Kaiming He, and Jian Sun. "Convolutional feature masking for joint object and stuff segmentation." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.

[9] Hariharan, Bharath, et al. "Simultaneous detection and segmentation." Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VII 13. Springer International Publishing, 2014.

[10] Huang, Zilong, et al. "Ccnet: Criss-cross attention for semantic segmentation." Proceedings of the IEEE/CVF international conference on computer vision. 2019.

[11] Long, Jonathan, Evan Shelhamer, and Trevor Darrell. "Fully convolutional networks for semantic segmentation." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.

[12] Strudel, Robin, et al. "Segmenter: Transformer for semantic segmentation." Proceedings of the IEEE/CVF international conference on computer vision. 2021.

[13] Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).

[14] Xie, Enze, et al. "SegFormer: Simple and efficient design for semantic segmentation with transformers." Advances in Neural Information Processing Systems 34 (2021): 12077-12090.

[15] He, Kaiming, et al. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[16] Zhao, Hengshuang, et al. "Pyramid scene parsing network." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.

[17] Liu, Chenxi, et al. "Auto-deeplab: Hierarchical neural architecture search for semantic image segmentation." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019.

[18] Wang, Jingdong, et al. "Deep high-resolution representation learning for visual recognition." IEEE transactions on pattern analysis and machine intelligence 43.10 (2020): 3349-3364.

[19] Wang, Wenhai, et al. "Internimage: Exploring large-scale vision foundation models with deformable convolutions." arXiv preprint arXiv:2211.05778 (2022).