IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Traitement du Signal
Vol. 40, No. 1, February, 2023, pp. 297-307
Journal homepage: http://iieta.org/journals/ts

Fusion of Ensembled UNET and Ensembled FPN for Semantic Segmentation

Kavitha Sundarrajan[1*], Baskaran Kuttva Rajendran[2], Dhanapriya Balasubramanian[1]

[1] Department of Information Technology, Kumaraguru College of Technology, Coimbatore 641049, India
[2] Department of Computer Science and Engineering, Kumaraguru College of Technology, Coimbatore 641049, India

Corresponding Author Email: kavitha.s.it@kct.ac.in

https://doi.org/10.18280/ts.400129

Received: 15 November 2022
Accepted: 6 February 2023

Keywords:
UNET, FPN, pre-trained model, F1 score, intersection over union, semantic segmentation, ensembling

ABSTRACT

Image segmentation is an annotation method used to gain a deeper understanding of the images. Semantic segmentation involves constructing a pixel-by-pixel mask of an image by training a neural network. The accuracy of the semantic segmentation algorithms can be improved by eliminating background noise, and computational efficiency can be improved by using the pre-trained networks. This paper proposes a new architecture that ensemble inceptionV3, DenseNet, Resnet34 in the encoder part of UNET and ensemble inceptionV3, Resnet34, and VGG16 in the encoder part of FPN. The ensemble results are fused based on the weighted average and the predictions of the pixels are made on the fused features to perform semantic segmentation. The proposed architecture is implemented on Oxford-IIIT Pet Dataset, created by the visual geometry group, and on the SD saliency 900 dataset. The F1 score, IOU score, and Loss are used to evaluate segmentation model results. The results of the study show that the proposed architecture formed by the fusion of ensembled architectures is more accurate and efficient in segmenting oxford-IIIT pet dataset with the IoU score of 98.68% and segmenting the SD saliency 900 dataset with the IoU score of 66.78%.

1. INTRODUCTION

The annotation of digital images is one of the most essential tasks in computer vision, as this is the process that allows systems to gain deeper insights into digital images. Adding an annotation to images is essential for the machines to properly identify and interpret the objects. The annotation of images involves labeling the objects that a model is expected to recognize in them. The images, along with tags representing the objects in them, are used to train the system, enabling it to recognize the objects in the new image. To enable computers to learn digital images, several types of annotation are available. Below is a list of the different annotation techniques available in computer vision:

·Image Classification: It is a form of annotation in which an image is analyzed, and based on the analysis; the image is classified into a specific category.

·Object Detection: It is a technique that enables computers to recognize objects in an image.

·Segmentation: Every pixel in an image is classified into a specific class through the segmentation process.

Segmentation is the advanced form of annotation. There are different types of segmentation, including semantic segmentation, panoptic segmentation, and instance segmentation. Semantic image segmentation is the process of classifying each pixel in the image. Semantic segmentation involves assigning the same values to pixels corresponding to a particular class. Multiple objects of the same class are treated as a single entity in semantic segmentation. Semantic segmentation is used to understand the presence, location, and sometimes, the size and shape of objects present in the image. The semantic segmentation process simplifies the analysis of the images by aiding in precise object recognition. Semantic segmentation is the most efficient one because it segments images more precisely and, therefore, makes it easier for systems to identify the desired objects in images.

Nizam et al. [1] have implemented Faster RCNN for the segmentation process. The meliponine family consists of bees that are very small in size. The meliponine family is very difficult to identify in the natural environment. Meliponine image is segmented from its background using Fast R-CNN, an object detection method. As a result, faster RCNN produces 74% accuracy, which is a reasonable level of accuracy.

Another promising architecture used for semantic segmentation is FAST FCN (Fast Fully Convolutional Neural Network). The satellite images are classified using FCN [2] and all the images are grouped into the classes of water, farmland, and field. Thus, by implementing FAST FCN, the researchers were able to achieve an accuracy score of 0.93 and a precision score of 0.99.

The iris pattern is obscured by eyelids, eyelashes, and reflections, which can lead to segmentation errors because the iris lies within a small, damp, and dynamic area. Researchers are increasingly using convolutional neural networks (CNNs) to improve the accuracy of existing iris segmentation techniques. Lozej et al. [3] highlights several deep learning models implemented in UNET and applied to the CASIA dataset to provide precise results.

Mask RCNN algorithms [4] have been used to detect and segment tooth forms which help in achieving reasonable pixel accuracy. Researchers are also exploring the underwater environment. Semantic image segmentation is used to explore objects underwater [5]. Underwater fish are segmented accurately based on employing UNET as the semantic

segmentation framework. The architecture proposed, thus resulted in an IOU score of 0.8583.

To speed up and simplify the detection of liver cancer, this research intends to create a UNET architecture for segmenting the liver and tumor from abdominal CT scan pictures. In this work, the number of convolutional filters per block was decreased in a contracting path to improve the accuracy of the model [6]. Segmenting blood vessels has become very popular among researchers. This research proposes a novel approach for segmenting blood arteries using U-net Convolutional Neural Networks (CNNs). The suggested method is more precise, focused, and sensitive when compared to existing methods [7].

The detection and classification of tumors can benefit substantially from the use of medical pictures. Brain tumors are segmented into multiple classes using cascade CNN and distance wise attention mechanism. A multi-modality brain MRI image dataset called BRATS-2018 is utilized to train the suggested architecture. The proposed architecture performed better than other architectures, scoring 87.45 on the dice score [8].

For the segmentation three approaches are suggested: automatic region growth, fully automatic mask selection-based active contour techniques, and mass segmentation dependent on iterative active contour [9]. The experiment proves that the automatic mass segmentation based on the active contour technique results in segmentation with high accuracy. On the Cityscapes Dataset, the suggested approach delivers Panoptic Segmentation. Utilizing DeepLabV3 cutting-edge's encoder-decoder architecture, MobileNet-V2 has been customized and optimized. The proposed method also makes use of Atrous convolution and Spatial Pyramid Pooling to increase its accuracy and sturdiness. Very positive and encouraging results have been obtained, demonstrating the potential of the suggested strategy for rapid and accurate robust scene interpretation [10].

Kim et al. [11] proposed two models and three model ensembling to improve the performance of portrait semantic segmentation by employing soft voting and weighted soft voting. To evaluate the faculty performance of the Teaching assistant evaluation dataset authors employed stacking and voting ensemble methods to improve the performance of the classifier [12].

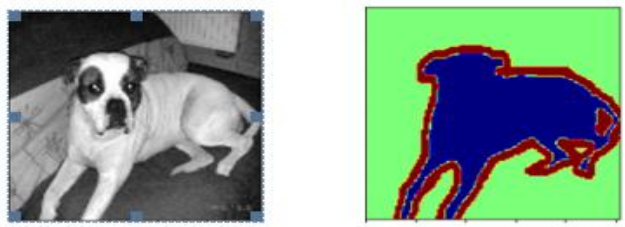
A multi-task model that executes pixel-based defect segmentation and severity estimate of the flaws in a single two-branch network was proposed [13]. First, two single-task models were developed and trained to perform the fault segmentation and severity estimation tasks independently. Then, they contrasted this to a multi-task model that executes the two tasks at hand simultaneously. Both segmentation tasks increased by 2.5% and 3% mIoU when the tasks were combined into a single model. The author presented a joint attention-guided feature fusion network (JAFFNet) based on the encoder-decoder network for the saliency identification of surface faults. The JAFF module gains the ability to highlight faulty features and reduce background noise during feature fusion for finding low-contrast faults [14].

In the existing literature, the fusion of heterogenous ensembled architecture is not carried out/implemented. In this paper, we propose a novel approach that fuses the ensembled UNET architecture with the ensembled FPN architecture. Ensembled UNET and ensembled FPN is formed by ensembling the three best pretrained deep learning models. The main goal of the proposed architecture is to improve the performance of semantic segmentation. By using the proposed architecture, a very high IOU has been achieved, and dice loss is very minimal when compared with the existing state of art methods in the literature.

2. DATASET

The Oxford-IIIT Pet Dataset, created by the Visual Geometry Group consists of 37 categories of pets [15]. Each category contains approximately 100 images. In addition to species and breed name each image in the dataset includes Trimap. Figure 1 Shows a sample image along with Trimap. A Trimap is a way of segmenting the foreground and background of an image pixel by pixel. Trimaps are essentially labels that are attached to each pixel. Pixels in images can be categorized into three subcategories:

- Class 1: Pixels associated with the pet.
- Class 2: Pixels that encircle the pet. (Borders).
- Class 3: The pixels surrounding the pet.



(a) (b)

Figure 1. Sample image from oxford-IIIT pet dataset along with trimap a) Sample image b) Trimap

SD saliency 900 Dataset consists of 3 categories (Patches, Scratches, Inclusion) of steel surface defect [16]. Each category contains approximately 300 images. In addition to the defect image, each image in the dataset includes a segmentation mask. Figure 2 shows a sample image along with Mask. The mask image has 189 unique pixel values.

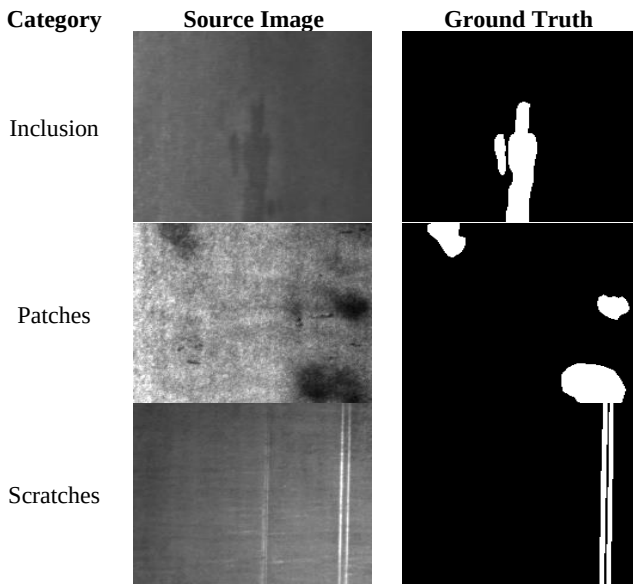


Figure 2. Sample category wise image from SD saliency 900 dataset along with the ground truth

3. METHODOLOGY

The proposed methodology as shown in Figure 3. fuses the ensembled UNET and ensembled FPN. First, the encoder of the UNET architecture is replaced by a single pretrained network (VGG, INCEPTION, RESNET, DENSENET, MOBILENET). Each model is evaluated with the IOU score, F1 score, and Loss function. We select the best three models (InceptionV3, DenseNet, and ResNet) based on their evaluation metrics. Second, the encoder of the FPN architecture is replaced by a single pretrained network (VGG, INCEPTION, RESNET, DENSENET, MOBILENET). Each model is evaluated with the IOU score, F1 score, and Loss function. We select the best three models (VGG16, InceptionV3, and ResNet) based on their evaluation metrics. The selected models are ensembled to form the segmented mask. To improve the performance of segmentation, the ensembled models are fused. The model results with the best IOU score of 98.68%.

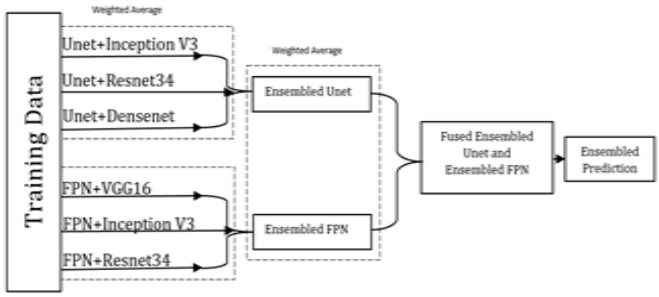


Figure 3. Proposed methodology

3.1 Segmentation using UNET and FPN

UNET is an architecture designed to facilitate the semantic segmentation process. UNET architecture consists of deep learning layers such as convolutional and maximum pooling layers, which are organized in such a way that it results in the segmentation process. Two paths are involved in the UNET architecture. The path on the left is known as the contracting path or encoder and the path on the right is known as the expansion path or decoder. In the left path, the encoder will perform the Maxpool Downsampling to extract the features. In the right path, the decoder will perform upsampling and concatenation to reproduce the image.

A feature extractor known as a Feature Pyramid Network or FPN generates proportionally scaled feature maps at several levels in a completely convolutional manner from a single-scale image of any size. A Top-down and a bottom-up pathway are used to build the pyramid. The feedforward computation of the backbone ConvNet's bottom-up pathway computes a feature hierarchy made up of feature maps at various scales with a scaling step of two. One pyramid level is established for each stage in the feature pyramid. As a reference set of feature maps, the output of each stage's final layer is used. We employ the feature activations produced by each stage's final residual block for ResNet.

The top-down pathway creates the illusion of greater resolution features. Through lateral connections, these features are then improved with features from the bottom-up pathway. Each lateral link combines feature maps from the top-down and bottom-up pathways that are the same spatial size. The bottom-up feature map has lower-level semantics, but because it was subsampled less frequently, its activations are more precisely localized.

3.1.1 Basic operations in UNET and FPN

Operations in UNET

Convolution:

In a convolution operation, features of the input image are extracted. It is the most essential and basic operation conducted in the UNET model.

$$(N\times N)*(F\times F)=(N-F+1)*(N-F+1) \quad (1)$$

where, N×N=Size of image and F×F=Size of filter.

Pooling:

Pooling is used to reduce the size of the feature maps thus making the computation faster.

The first step in the segmentation process is to resize the images to the standard size, so the images in the two datasets have been resized to 128*128. All the images are colored images; hence they include 3 channels, resulting in the image shape being 128*128*3.

In UNET architecture, the image is given to the input layer, and to this input image a feature space of 16 is added, resulting in a 128*128*16 result. Convolution is carried out with 3*3 kernel size and padding is modulated as same, which means additional pixels are added on the edges, making the output image the same size as the input. Following this operation is a max pooling operation, upsampling and a few convolutions are performed.

Operations in FPN

Botton-Up pathway:

ResNet is used to build the bottom-up pathway in this method. It is made up of numerous convolution modules (convi for I=1 to 5), each with numerous convolution layers. The spatial dimension is halved as we ascend (i.e. double the stride). Each convolution module's output is tagged before being used in the top-down pathway.

Top-down pathway:

We use the nearest neighbor upsampling to upsample the previous layer by 2 as we proceed down the top-down path. In the bottom-up pathway, we once again apply a 1×1 convolution to the relevant feature maps. We then add them element by element. To all merged layers, a 3×3 convolution is applied. When combined with the up-sampled layer, this filter minimizes the aliasing effect.

3.2 Segmentation carried out by replacing the encoder with the single pre-trained model in UNET and FPN

UNET architecture has two paths, one being the contracting path or encoder, while the other is the expansion path or decoder. The encoder part of UNET architecture can be replaced with any pre-trained model. In the UNET architecture, "backbone" refers to the pretrained model to be used as the encoder. This pre-trained network as a backbone facilitates the use of pre-trained weights (example: ImageNet). The decoder part can be built up programmatically based on the encoder part. FPN architecture has two paths, one being the Bottom or encoder, while the other is the Top-down or decoder. The encoder part of UNET and FPN architecture can be replaced with any pre-trained model. In the UNET and FPN

architecture, "backbone" refers to the pretrained model to be used as the encoder. This pre-trained network as a backbone facilitates the use of pre-trained weights (example: ImageNet). The decoder part can be built up programmatically based on the encoder part. The Segmentation Model library contains UNET and FPN architectures as well as several backbones. All backbones have weights trained on the IMAGENET database. This library constructs the decoder automatically based on the backbone used.

A specific pre-trained model is imported based on the backbone used. Data is pre-processed using the get preprocessing module based on the backbone used as the encoder block. As semantic segmentation is simply a multiclass classification, SoftMax activation is used. As the activation is SoftMax, the output is the probabilities for each channel. The modified UNET and FPN model is compiled using Adam optimizer along with a learning rate of 0. 0001. The loss function used in the compilation is a combination of dice loss and focal loss. These two losses are used for semantic segmentation purposes to treat unbalanced data. The metrics used in the compilation process are the F1 score and the IOU score. In this way, the compiled model is then trained in batches of 8 for about 50 epochs, making it more effective, so that when a test image is provided, every pixel is correctly classified.

In this paper, the UNET's and FPN's encoder block is replaced with five pre-trained networks VGG16, RESNET, INCEPTIONV3, DENSENET, and MOBILENETV2. Following are the reasons behind choosing each of these backbones:

• VGG16 - It reduces the number of parameters in convolutional layers thus improving the training time.

• RESNET - It is used as a backbone because it eliminates the vanishing gradient problem.

• INCEPTIONV3 - Choosing a fixed kernel size for image analysis can be difficult since features may vary in size. For global features, larger kernels are preferred over a larger area of the image. Area-specific features are best detected with smaller kernels. InceptionV3 helps in achieving this by going wider with layers. Within the same layer, kernels of different sizes are used.

• DENSENET-DenseNet reduces the number of parameters and thus motivates the reuse of features.

• MOBILENETV2-MobilenetV2 has very small computational complexity and is one of the fastest models, resulting in very high accuracy.

3.3 Semantic segmentation via an ensemble of pre-trained networks in UNET and FPN

Figures 4 and 5 show the ensembling of three pretrained networks in the UNET and FPN architectures. Based on the IOU score of single pretrained networks, the best three models were selected for ensembling. The UNET architecture uses three pre-trained networks (InceptionV3, ResNet, and DenseNet) for ensembling.

The FPN architecture uses three pre-trained networks (VGG16, InceptionV3, and ResNet) for ensembling. These pre-trained models are trained independently and then predictions from these pre-trained models are taken as outputs.

The three outputs, which are mainly probabilities based on replacing the encoder part of UNET and FPN with three pre-trained networks individually, are stored in a list and converted to an array. A set of weights is assigned, and the probabilities of all three pretrained models are multiplied by those weights, and then the summed probabilities are calculated.

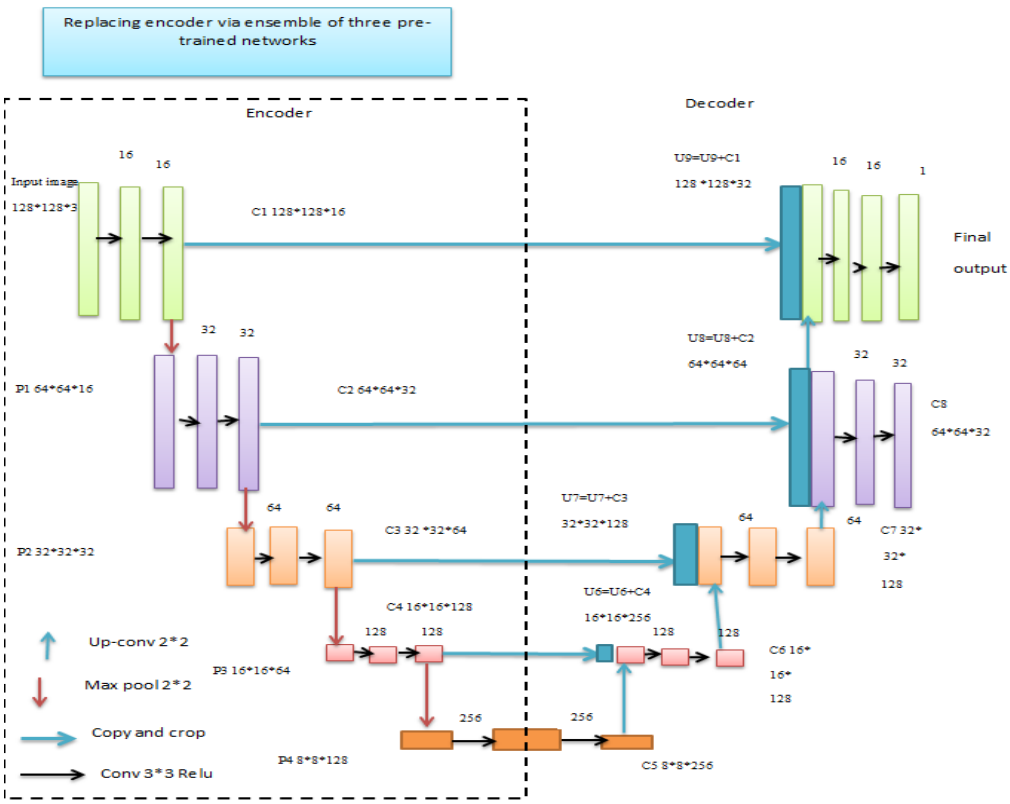


Figure 4. Ensemble of pre-trained networks in UNET

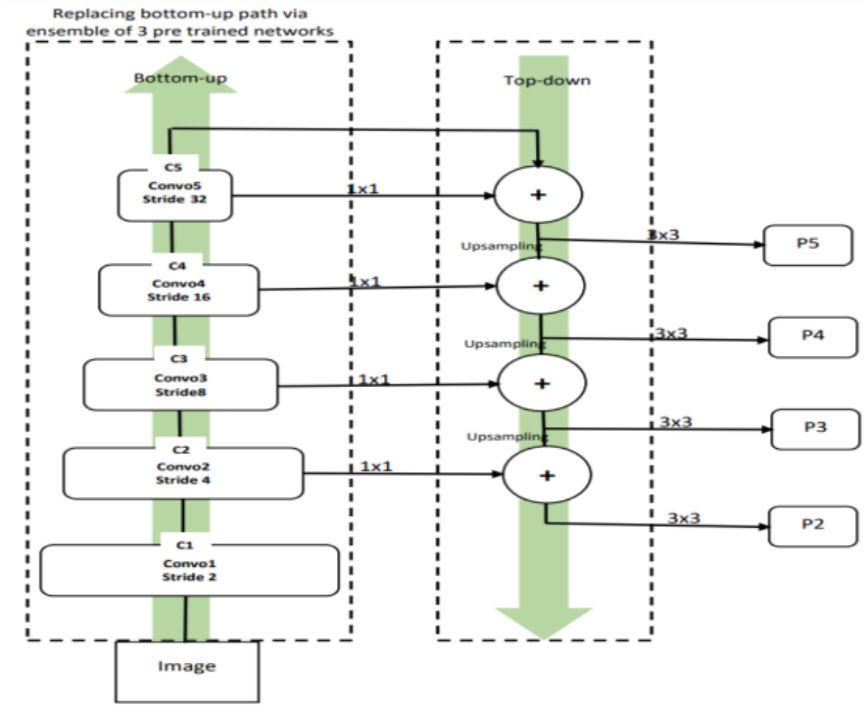


Figure 5. Ensemble of pre-trained networks in FPN

Segmentation models are a pre-defined library that has UNET architecture and FPN architecture that supports multiple backbones. Based on the backbone specified in UNET and FPN architecture, the segmentation models library will define the decoder. As these backbones were trained on huge datasets such as IMAGENET, defining the backbone results in defining the weights of the pre-trained networks.

As part of the proposed architecture, four pre-trained networks will be utilized as encoders – VGG16, InceptionV3, ResNet, and DenseNet. The reason behind using these pre-trained networks is stated below:

·VGG16

The VGG-16 model is trained based on the extracted features as input data. This relatively small model on low-dimensional data does well. The layers of the model are 13 convolutional layers, 5 pooling layers, and 3 dense layers.

·ResNet

As layers get increased, the error rate also increases but with the use of ResNet, the percentage of error taken during training gets reduced. Using ResNet, the vanishing gradient issue gets solved using skip connections in ResNet.

·DenseNet

DenseNet is used for its ability to reduce the number of parameters, as well as the need to memorize duplicate feature maps. A DenseNet is split into Dense Blocks, where the dimension of the feature maps is the same within a block, but the number of filters differs between them. The output feature maps of the layer are not added to the incoming feature maps by DenseNets but are concatenated. Hence the equation is represented as:

$$x_l = H_l[x_0, x_1, \ldots\ldots x_{l-1}] \tag{2}$$

·InceptionV3

InceptionV3 architecture is used due to its huge performance gains. Using different convolutional filter sizes, the Inception network can extract features from input data at varying scales. Compared to other architectures, it is less expensive. Below are the steps that are used in training these networks independently as an encoder in UNET.

Images and masks are resized to the standard size of 128*128. Oxford pet dataset contains three classes, every pixel in the mask can be categorized into any one of the three classes. The resized images and resized masks are therefore stored in a list, which is then converted to an array. To analyze the images in the dataset, images are read in RGB format, so they have 3 channels, while masks are read in rescale format. The class values are converted to numeric values (0, 1, 2) using a label encoder. In SD saliency 900 mask dataset pixel in the mask can be categorized into any one of the 189 classes. To analyze the images in the dataset, images are read in RGB format, so they have 3 channels, while masks are read in rescale format. The class values are converted to numeric values (from 0 to 188) using a label encoder. The labels are then one hot encoded using the to_categorical method.

The parameters which are essential for the model construction are defined as follows:

·Activation

SoftMax activation is used as semantic segmentation a form of multiclass classification.

$$\sigma(Z) = (e^Z)/\sum_{j=1}^{K} e^{zj} \tag{3}$$

where, σ denotes the softmax activation.

·Loss function

The summation of dice and focal loss is used as these losses are best for the semantic segmentation process.

·Optimiser

Adam optimizer is used and has a faster computation time and also requires fewer parameters for tuning.

$$w_t = w_{t-1} - n(m_t)/(\sqrt{V_t})+\in \tag{4}$$

where, *n*=learning rate.

·Metrics: F1 score and IOU score.

Pre-processing is done using the get preprocessing module for input images depending on the backbone used. The compilation of the model is then based on the specified backbone, optimizer, and metrics defined. Afterward, the model is trained in batches of 8 for about 50 epochs. The above procedure is thus repeated by defining three pre-trained networks (ResNet, InceptionV3, and DenseNet) as UNET'S backbone and three pre-trained networks (ResNet, InceptionV3, and VGG16) as FPN backbone.

The two outputs, which are mainly probabilities based on replacing the encoder part of UNET with three pre-trained networks individually, and probabilities based on replacing the encoder part of FPN with three pre-trained networks individually are stored in a list and converted to an array. A set of weights is assigned, and the probabilities of all two ensembled models are multiplied by those weights, and then the summed probabilities are calculated. Based on the fused output, the predictions are made on the test image thus resulting in precise and more accurate results.

4. RESULTS AND DISCUSSIONS

4.1 Metrics used

·IOU Score

The intersection over union (IOU) metric, or the Jaccard index, is a way to measure how much the target mask overlaps with our prediction output. In other words, IOU is defined as the number of pixels shared between the target and prediction masks divided by the total number of pixels.

$$IOU = \frac{target \cap prediction}{-target \cup prediction} \quad (5)$$

·F1 score

The balance between precision and recall is called the "F1 score". F1 score varies from 0 to 1.

$$F1\ Score = (2 * TP)/(2 * TP + FP + FN) \quad (6)$$

4.2 Loss function

·Dice loss

Dice loss is mainly used to address data imbalance problems. It is mainly used to calculate the similarity between two images.

$$Diceloss = 1 - Dice\ coefficient \quad (7)$$

where, the dice coefficient is the measure of overlap between two masks.

·Focal loss

Focal loss is also used to address the data imbalance problem. It is defined as the addition of modulating terms to the cross-entropy loss.

$$Focal\ loss = -\alpha t(1 - pt)\gamma \log(pt) \quad (8)$$

·Total loss

The total loss is the summation of dice loss and focal loss.

$$Total\ loss = Dice\ loss + Focal\ loss \quad (9)$$

4.3 Results of single models in UNET and FPN

Below is Table 1. which gives the metrics values of all the architectures obtained by replacing the encoder part of the UNET with the single pretrained networks. The top three IOU scores are given in bold.

Table 1. IOU score, F1 score, and loss obtained by replacing backbones of UNET architecture

Semantic Segmentation	IOU	F1 Score	Dice loss
OXFORD – IIITPET DATASET			
Unet	0.333	0.464	0.535
Unet+VGG16	0.894	0.942	0.076
Unet+Inception V3	**0.916**	0.915	0.059
Unet+Resnet34	**0.910**	0.951	0.063
Unet+Densenet	**0.915**	0.953	0.059
Unet+MobilenetV2	0.896	0.923	0.066
SD SALIENCY 900 DATASET			
Unet	0.5937	0.5960	0.4935
Unet+VGG16	0.6063	0.6137	0.4733
Unet+Inception V3	**0.6196**	0.6310	0.4595
Unet+Resnet34	**0.6187**	0.6298	0.4604
Unet+Densenet	**0.6195**	0.6306	0.4612
Unet+MobilenetV2	0.6187	0.6301	0.4592

Table 2 gives the metrics values of all the architectures obtained by replacing the encoder part of the FPN with the single pretrained networks on the two datasets. The top three IOU scores are given in bold.

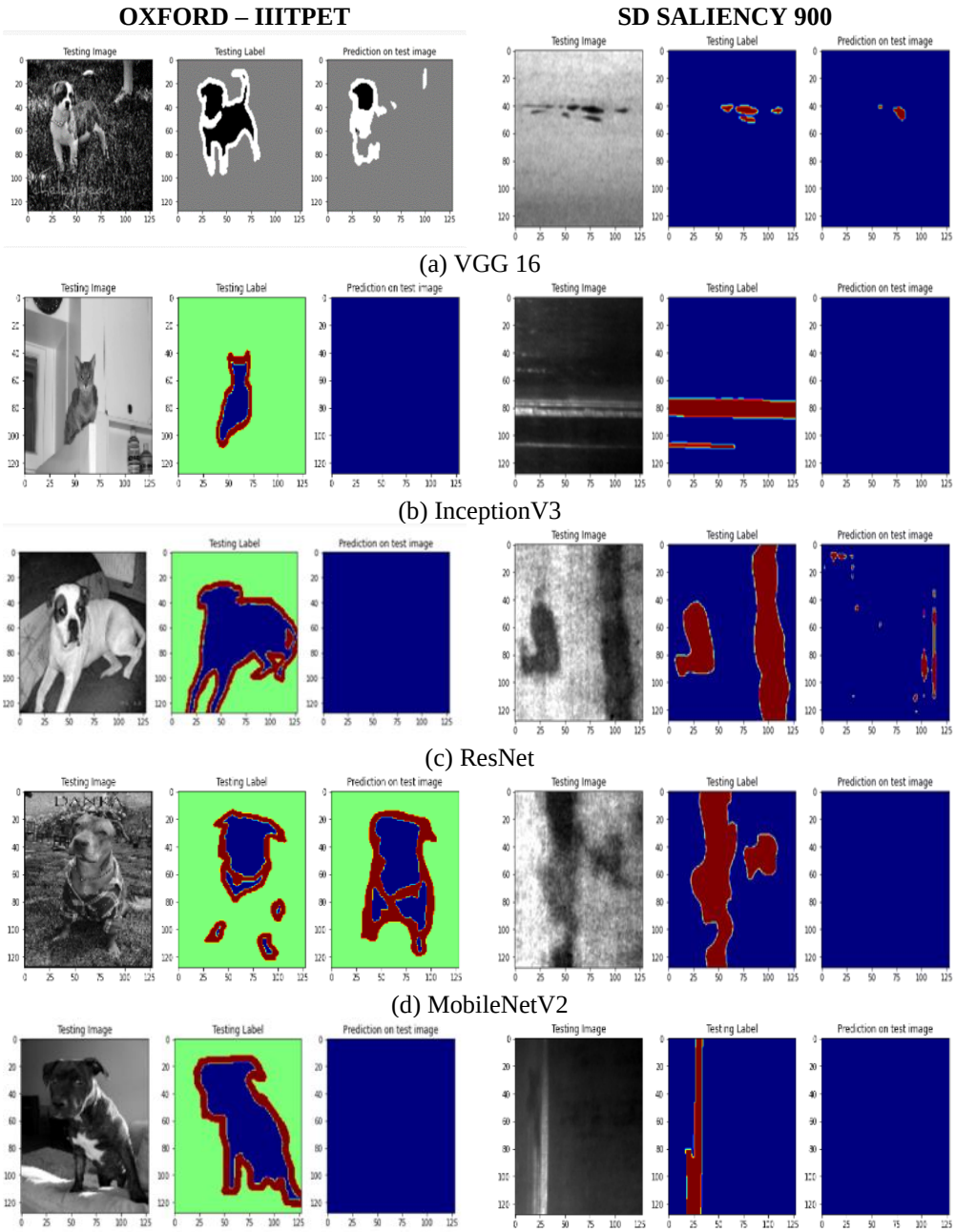
Table 2. IOU score, F1 score, and loss obtained by replacing Backbones of FPN architecture

Semantic Segmentation	IOU	F1 Score	Dice loss
OXFORD – IIITPET DATASET			
FPN+VGG16	**0.9591**	0.9788	0.0268
FPN+Inception V3	**0.9329**	0.9644	0.0444
FPN+Resnet34	**0.9410**	0.9688	0.0388
FPN+Densenet	0.9273	0.9612	0.0481
FPN+MobilenetV2	0.9079	0.9498	0.0626
SD SALIENCY 900 DATASET			
FPN+VGG16	**0.6552**	0.6851	0.3919
FPN+Inception V3	**0.6581**	0.6905	0.3837
FPN+Resnet34	**0.6531**	0.6866	0.3856
FPN+Densenet	0.6531	0.6848	0.3866
FPN+MobilenetV2	0.6395	0.6676	0.3689

From the Tables 1 and 2, The three best IOU scores are selected, and the models were used for ensembling in the encoder part of UNET and FPN architectures.

The predictions obtained by replacing the encoder part of UNET architecture with the above-mentioned pre-trained models on the oxford-IIIT pet dataset and SD Saliency 900 individually are shown in Figure 6.

The predictions obtained by replacing the encoder part of FPN architecture with the above-mentioned pre-trained models individually are shown in Figure 7.



(a) VGG 16

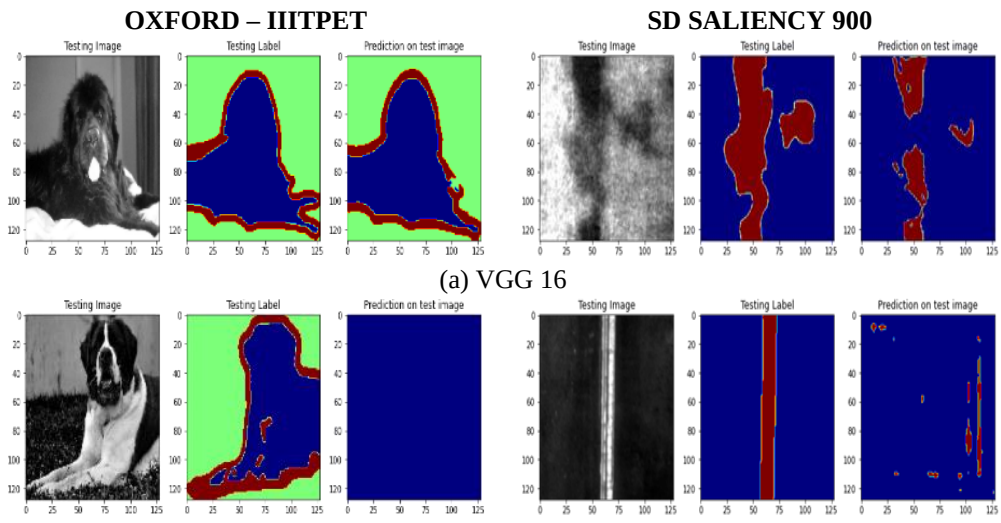
(b) InceptionV3

(c) ResNet

(d) MobileNetV2

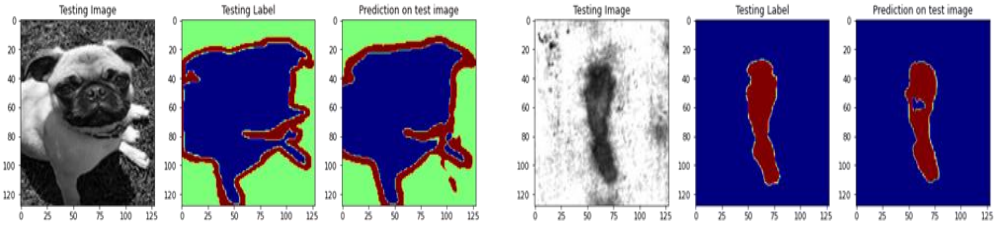
(e) DenseNet

Figure 6. Predictions on test image by replacing encoder in UNET with single pre-trained networks

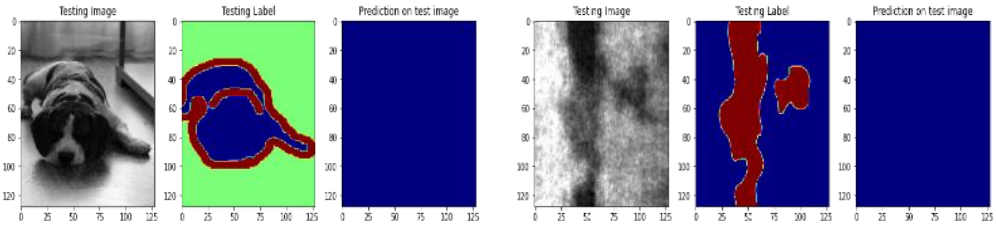


(a) VGG 16

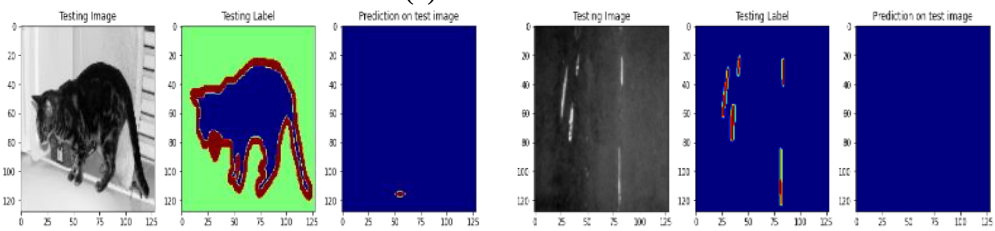
(b) InceptionV3



(c) ResNet



(d) MobileNetV2



(e) DenseNet

Figure 7. Predictions on test image by replacing encoder in FPN with single pre-trained networks

4.4 Results of ensembled UNET and ensembled FPN

Table 3 gives the metrics values of the architectures obtained by the ensembling of best models in the UNET and FPN. Ensembling in the UNET and FPN is done by considering the weighted average of the predicted class for each pixel value. Ensembling the best IoU score models improve the IoU score, and F1 score and reduce the Dice loss.

Table 3. IOU score, F1 score, and loss obtained by ensembled UNET and ensembled FPN architecture

Semantic Segmentation	IOU	F1 Score	Dice loss
OXFORD – IIIT PET DATASET			
Ensembled Unet (Inception + Resnet + DenseNet)	0.948	0.956	0.057
Ensembled FPN (Inception V3 + Resnet + VGG16)	0.974	0.987	0.017
SD SALIENCY 900 DATASET			
Ensembled Unet (Inception + Resnet + DenseNet)	0.620	0.636	0.447
Ensembled FPN (InceptionV3 + Resnet + VGG16)	0.664	0.699	0.357

A graph that is drawn between the training and validation loss for the Ensembled UNET and Ensembled FPN is shown in Figure 8 and Figure 10. A graph that is drawn between the training and validation IOU for the Ensembled UNET and Ensembled FPN obtained on the Oxford-IIIT pet dataset is shown in Figure 9 and Figure 11.

Figures 12 and 13 Shows the predicted output obtained by the ensembled UNET and Ensembled FPN.

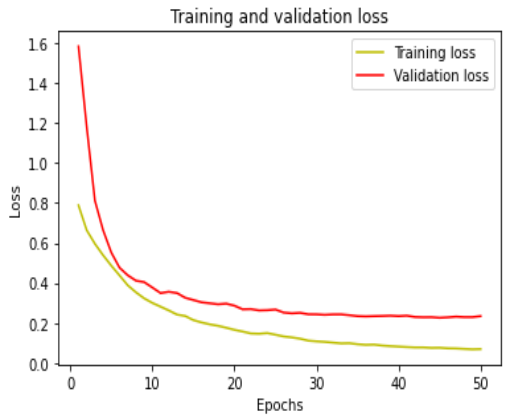


Figure 8. Training and validation loss obtained by Ensembled UNET

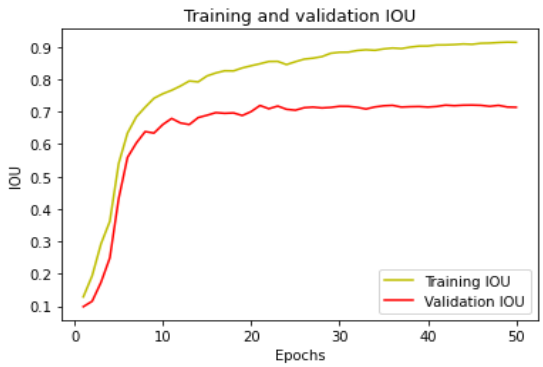


Figure 9. Training and validation IOU obtained by Ensembled UNET Architecture

Figure 10. Training and validation loss obtained by Ensembled FPN Architecture

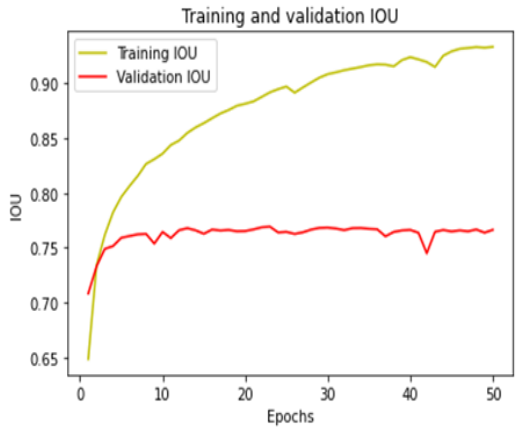


Figure 11. Training and validation IOU obtained by Ensembled FPN Architecture

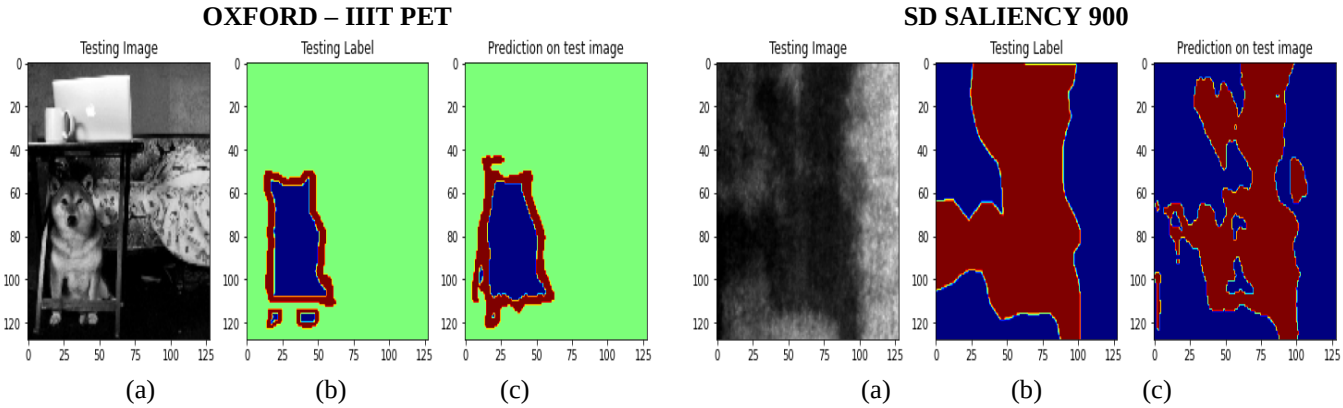


Figure 12. a) Sample image b) Ground truth label c) Prediction made by Ensembled UNET

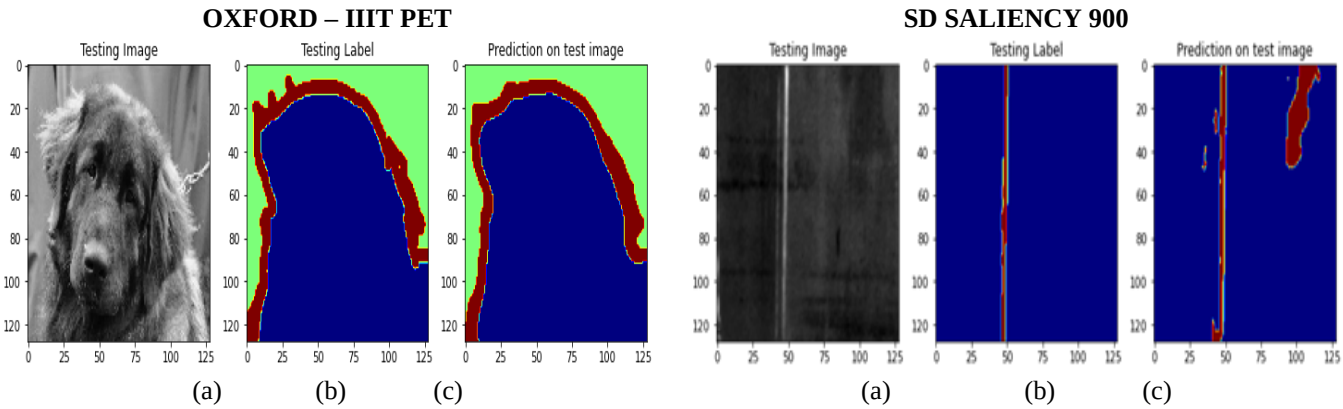


Figure 13. a) Sample image b) Ground truth label c) Prediction made by Ensembled FPN

4.5 Results of fused ensembled UNET and ensembled FPN

The weighted average of the predicted class values of each pixel value is calculated by considering the output of Ensembled UNET and Ensembled FPN. The proposed model achieves the high IOU of 98.68 because of the ensembling of the models with high IoU and low dice loss in the UNET and FPN architecture. Fusion of the ensembled architecture improves the IoU score. Table 4 gives the Evaluation metric values of the architectures obtained by the fusion of ensembled UNET and FPN models.

A graph that is drawn between the training and validation loss for the proposed architecture is shown in Figure 14. From the graph, it is evident that as the epoch increases the loss obtained is also minimized thus making the semantic segmentation process more effective in the oxford-IIIT pet dataset.

Table 4. IOU score, F1 score, and loss obtained by fusion of ensembled UNET and ensembled FPN architecture

Semantic Segmentation	IOU	F1 Score	Dice loss
OXFORD – IIIT PET DATASET			
Ensembled UNET+ Ensembled FPN	0.9868	0.9978	0.0157
SD SALIENCY 900 DATASET			
Ensembled UNET+ Ensembled FPN	0.6678	0.7078	0.3557

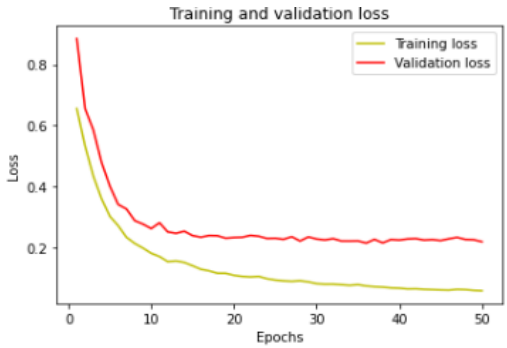


Figure 14. Training and validation loss obtained by fusion of ensembled UNET and ensembled FPN architecture

A graph that is drawn between the training and validation IOU score for the proposed architecture is shown in Figure 15. From the graph, it is evident that as the epoch increases the IOU score is also increased thus making the segmentation process more precise and accurate.

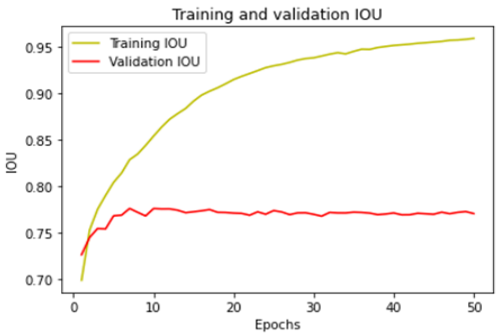


Figure 15. Training and validation IOU obtained by fusion of ensembled UNET and ensembled FPN architecture

Figure 16 shows the comparative study of Semantic segmentation models based on the evaluation metrics and it is evident that the proposed model is more precise and accurate with the IOU score of 98.68% on the oxford-IIIT pet dataset.

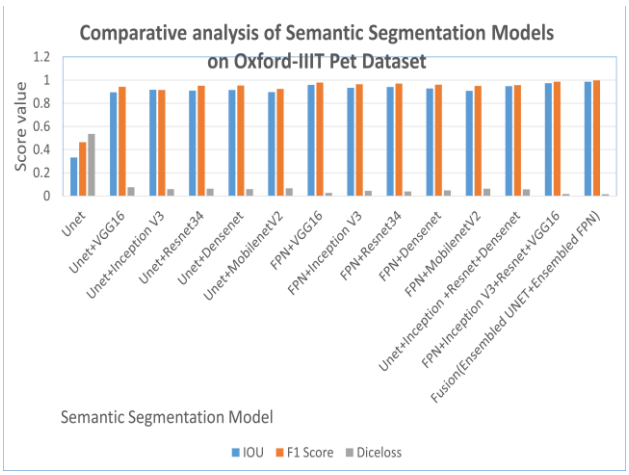


Figure 16. Comparative analysis of semantic segmentation models based on IOU, F1 score, and dice loss

Figure 17 shows the comparative study of Semantic segmentation models based on the evaluation metrics and it is evident that the proposed model is more precise and accurate with the IOU score of 66.78% on the SD Saliency 900 dataset.

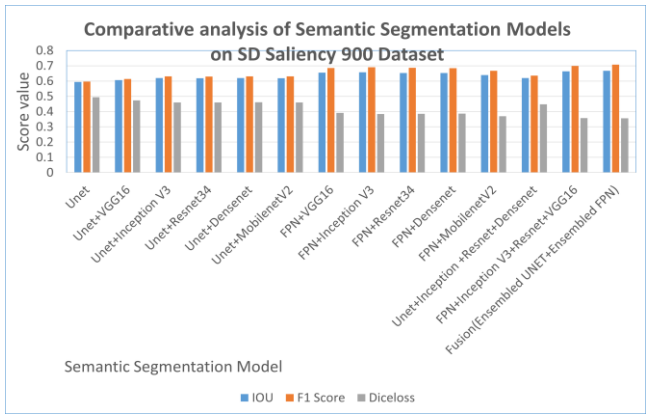


Figure 17. Comparative Analysis of Semantic segmentation models based on IOU, F1 score, and Dice loss

A sample test image is predicted using the proposed architecture as shown below. From Figures 18 and 19, it is evident that the proposed architecture segments image by classifying the pixels of images into specific classes more precisely.

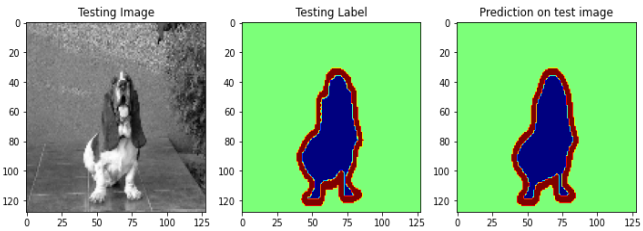


Figure 18. a) Sample image b) Ground truth label c) Prediction made by the proposed architecture

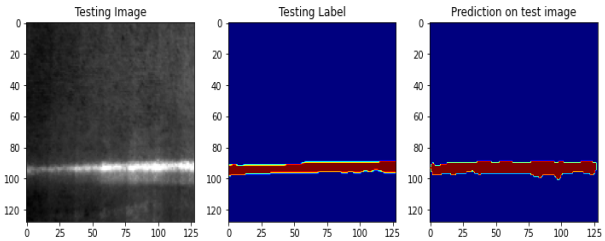


Figure 19. a) Sample image b) Ground truth label c) Prediction made by the proposed architecture

The IoU score of the proposed system is 66.78% on the SD saliency 900 dataset because the mask or the ground truth image has 189 unique pixel values. The proposed architecture works well in the oxford-IIIT pet dataset because of the trimap mask which has 3 unique pixel values.

5. CONCLUSIONS

The process of image segmentation is used in application software to analyze what is contained within images. The meaningful objects in an image are determined by image segmentation. Many applications require accurate and efficient image segmentation mechanisms to evaluate visual content and make inferences from it. Different networks are

proposed to improve the segmentation process. As part of this paper, we experimented with five pre-trained networks in the UNET and the FPN architectures. A new architecture was proposed that fuses the ensembled UNET and Ensembled FPN. The model was evaluated with two datasets and compared with the single pretrained networks used for semantic segmentation. The proposed fusion of ensembled network obtained the highest IoU score of 98.68 on the Oxford IIIT pet dataset as its mask has three unique pixel values while on the SD saliency 900 dataset IoU score is 66.78 as its mask has 189 unique pixel values.

REFERENCES

[1] Nizam, A., Mohd-Isa, W., Ali, A. (2020). Image segmentation of meliponine bee using mask-RCNN. International Journal of Engineering Trends and Technology, 17-21. https://doi.org/10.14445/22315381/cati2p203

[2] Singh, N., Nongmeikapam, K. (2022). Semantic segmentation of satellite images using deep-unet. Arabian Journal for Science and Engineering, 48: 1193-1205. https://doi.org/10.1007/s13369-022-06734-4

[3] Lozej, J., Meden, B., Struc, V., Peer, P. (2018). End-to-end iris segmentation using u-net. In 2018 IEEE International Work Conference on Bioinspired Intelligence (IWOBI), pp. 1-6. https://doi.org/10.1109/IWOBI.2018.8464213

[4] Zhu, G., Piao, Z., Kim, S.C. (2020). Tooth detection and segmentation with mask R-CNN. In 2020 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), pp. 070-072. https://doi.org/10.1109/ICAIIC48513.2020.9065216

[5] Nezla, N.A., Haridas, T.M., Supriya, M.H. (2021). Semantic segmentation of underwater images using unet architecture based deep convolutional encoder decoder model. In 2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS), pp. 28-33. https://doi.org/10.1109/ICACCS51430.2021.9441804

[6] Ayalew, Y.A., Fante, K.A., Mohammed, M.A. (2021). Modified U-Net for liver cancer segmentation from computed tomography images with a new class balancing method. BMC Biomedical Engineering, 3: 1-13. https://doi.org/10.1186/s42490-021-00050-y

[7] Sambyal, N., Saini, P., Syal, R., Gupta, V. (2020). Modified U-Net architecture for semantic segmentation of diabetic retinopathy images. Biocybernetics and Biomedical Engineering, 40(3): 1094-1109. https://doi.org/10.1016/j.bbe.2020.05.006

[8] Ranjbarzadeh, R., Kasgari, A.B., Ghoushchi, S.J., Anari, S., Naseri, M., Bendechache, M. (2021). Brain tumor segmentation based on deep learning and an attention mechanism using MRI multi-modalities brain images. Scientific Reports, 11: 10930. https://doi.org/10.1038/s41598-021-90428-8

[9] Yuvaraj, K., Ragupathy, U.S. (2022). Hybrid active contour mammographic mass segmentation and classification. Computer Systems Science and Engineering, 40(3): 823-834.

[10] Majid, A., Kausar, S., Tehsin, S., Jameel, A. (2022). A fast panoptic segmentation network for self-driving scene understanding. Computer Systems Science and Engineering, 43(1): 27-43.

[11] Kim, Y.W., Byun, Y.C., Krishna, A.V. (2021). Portrait segmentation using ensemble of heterogeneous deep-learning models. Entropy, 23(2): 197. https://doi.org/10.3390/e23020197

[12] Ahuja, R., Sharma, S.C. (2021). Stacking and voting ensemble methods fusion to evaluate instructor performance in higher education. International Journal of Information Technology, 13: 1721-1731. https://doi.org/10.1007/s41870-021-00729-4

[13] Neven, R., Goedemé, T. (2021). A multi-branch U-Net for steel surface defect type and severity segmentation. Metals, 11(6): 870. https://doi.org/10.3390/met11060870.

[14] Jiang, X., Yan, F., Lu, Y., et al. (2022). Joint attention-guided feature fusion network for saliency detection of surface defects. IEEE Transactions on Instrumentation and Measurement, 71: 1-12. https://doi.org/10.1109/TIM.2022.3218547

[15] The Oxford-IIIT Pet Dataset. https://www.kaggle.com/datasets/devdgohil/the-oxfordiiit-pet-dataset, accessed on 01-03-2022.

[16] SD-saliency-900. https://www.kaggle.com/datasets/alex000kim/sdsaliency900, accessed on 13-05-2022.