Chapter 5

Segmentation of Specular Reflection

Traditional image segmentation involves applying various image processing techniques like thresholding, edge detection, region growing or clustering algorithms. These techniques separate the different areas of the images based on the intensity, color, texture or spatial properties of the images. The main drawback of the traditional method is over-segmentation, which decreases the efficiency of the images, and also, these techniques are susceptible to noise. The deep learning segmentation model is used to reduce the extraction procedure and learn the complex features, resulting in a more accurate and robust segmentation process to overcome this challenge. However, the difficulty faced in the CNN is the annotated data and the large datasets for training the model. So, in this chapter, the CNN based segmentation model is used to segment the specular reflection on the smart colposcopy images in the automated process and to improve the accuracy of the segmentation

5.1 Overview of this Chapter

The CNN segmentation model is the critical concept used in computer vision to segment specific parts from the images. This chapter employs the CNN based segmentation model to predict the reflection regions from the colposcopy images to perform pixel-level segmentation using masked images. This chapter uses the binary masking to annotate the reflection region on the smart colposcopy images. The original and masked images are taken as the input to train the deep-learning segmentation models. The models like *"FCN, SegNet and U-Net"* models segment the SR pixel on smart colposcopy images. On analysis, the U-Net outperform the other segmentation model due to high segmentation accuracy with the minimum dataset. Since there are different versions of the U-Net model like *"U-Net++ and Residual U-Net"* models are compared to identify the suitable segmentation model for segmenting the glare region from smart colposcopy image. Based on the analysis, UNet++ model anticipate the SR pixels with higher accuracy. The U-Net++ has been improved to make the model more suitable for the of SR pixel segmentation. The overview for segmentation of SR using a segmentation model is shown in Figure 5.1.

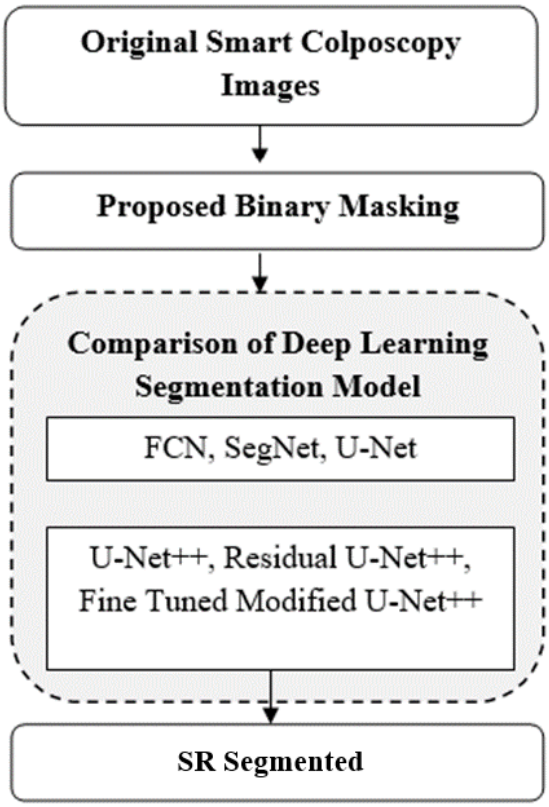


Figure 5.1. Workflow for segmentation SR Pixels

5.2. Binary Masking

An annotated dataset is an essential factor for training deep-learning segmentation models. The models are designed to segment specific regions within an image. Annotated data sets provide valuable ground truth labels or masks that accurately define the boundaries or pixel-level annotations of the areas of interest in the images. To make the accurate predictions, the CNN models require labelled data to learn patterns. Since the SR annotated dataset is unavailable here, the annotated dataset is created using the binary masking techniques using the proposed approach of intensity-based threshold method. Binary masking is used in various image processing application, where certain regions of the images are selectively masked or hidden. To label the reflection region in smart colposcopy images using binary masking, the following steps are involved:

Step 1: Iterating through each pixel of the image

- The loop for each x-coordinate starts from 0 to *(img_height - 1).*
- Inside the x-coordinate loop, a nested loop for each y-coordinate will start from 0 to *(img_width - 1).*

Step 2: RGB color Space of the current pixel at position (y, x) are converted to the XYZ color space

- RGB color values of the current pixel at position (*y, x)* are retrieved.
- The RGB color values of the images are remodeled to XYZ color space.

Step 3: The "current_pixel_intensity" variable is as range of the pixel (y, x) in the XYZ color space

- The "*current_pixel_intensity"* represents the pixel's intensity in the XYZ color space as represented in Algorithm 5.1.

Algorithm 5.1:

Input: *Image Dataset represented as a 2D array "pixel" with size (img_height, img_width)*

For *each x in range 0 to (img_height - 1)*
For *each y in range 0 to (img_width - 1)*

Convert the RGB color values of the pixel at position (y, x) to XYZ color space. Set current_pixel_intensity = value of "pixel" at position (y, x) in the XYZ color space
Set current_pixel = value of "pixel" at position (y, x)

If *current_pixel>= 200*
Set the value of "pixel" at position (y, x) to 1
Else if *current_pixel<= 200*
Set the value of "pixel" at position (y, x) to 0
End

Step 4: The current_pixel_intensity is checked and assigned the binary value to the corresponding pixel:

- If "*current_pixel_intensity"* is larger than or equal to 200, the pixel's value at position *(y, x)* is set as 1. It indicates that the pixel value falls in the SR region.
- If "*current_pixel_intensity"* is less than 200, the pixel's value at position *(y, x) is* set as 0. It indicates that the pixel value falls in the non-SR region.

Step 5: Steps 3 and 4 are repeated until all pixels of the images are converted to the binary-masked image.

- This algorithm generates the binary masked images, where the value 1 represents the reflection region in the colposcopy images, and the value 0 represents the non-reflection region of the images.

The SR identified on the images are masked and set as zero. In contrast, the other region of the non-specular reflection region is set as one, as shown in equation 5.1. These original images and masked binary images are considered as the input in the deep learning segmentation model to predict SR on the smart colposcopy images.

$$f(y,x) = \begin{cases} 1 = NonspecularReflection \\ 0 = SpecularReflection \end{cases} \quad (5.1)$$

The SR pixels are detected using the proposed approach of IRBM using XYZ color space and using these images the binary mask is generated, as demonstrated in Figure 5.2. Figure 5.2 (a) indicates the original images with SR. Figure 5.2 (b) indicates the cervical photo with SR pixel recognized with the method proposed. Figure 5.2 (c) illustrates the image masked produced from the method proposed. It is utilized as ground truth for segmentation of SR pixels on the smart colposcopy images.

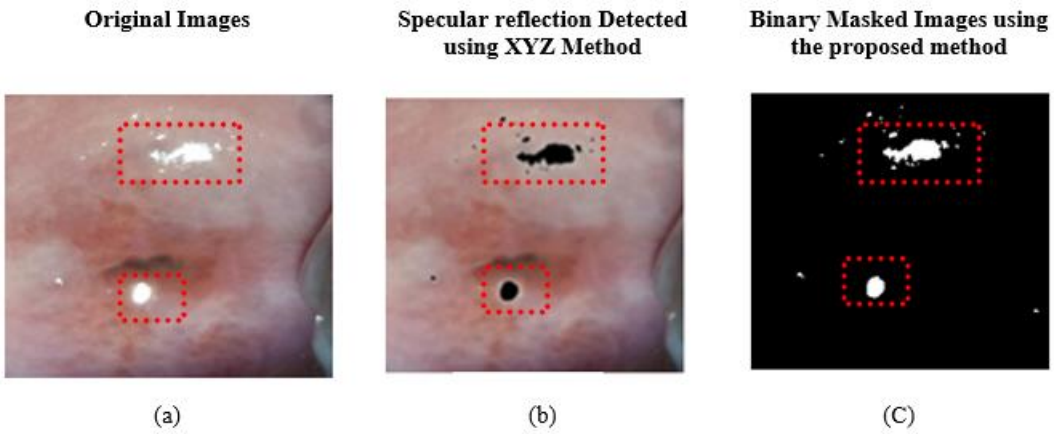


Figure 5.2 Creating Binary Masked images using the Proposed Threshold Method. (a) Original images with specular reflection (Reflection region marked as red). (b) Specular reflection was detected using the proposed threshold method on XYZ color space images. (c) Created the binary masked image using the proposed threshold method

5.3. CNN Based Segmentation Models

The CNN segmentation models utilize deep neural networks to automatically recognize and identifies features from images, enabling accurate segmentation results. This section discusses a CNN based deep learning architecture for segmenting specular reflections on smart colposcopy images. This section also explores key terms for constructing deep learning segmentation models. CNN segmentation models grasp the power of NN to recognize and identify features directly from images. It comprises of multiple layers, including convolutional layers that configure filters to acquire features from

input images and pooling layers that simplify the feature maps to capture spatial information efficiently. Binary masked images and original images which is resized are taken as the input images for the segmentation of SR pixels on cervix images. These images have a binary mask which specify the SR and non-SR on the images. The deep learning model aims to segment the specular reflections based on this binary mask accurately. Several key terms and concepts are essential to understand to construct effective deep learning segmentation models. Some of these terms include:

- **Convolution:** The process of applying a convolutional kernel/filter to an input image or feature map, outcoming in a transformed output feature map.
- **Convolutional Layer**: The layer in a CNN performs convolution operations on the input data, extracting local features.
- **Kernel/Filter**: A small matrix of weights used in convolution process to separate features from the input.
- **Stride:** The series of steps the convolutional kernel takes while moving through the input image or feature map.
- **Padding**: It adds additional pixels around the feature map to preserve spatial dimensions and avoid information loss during convolution.
- **Activation Function**: It is a non-linear function utilized to the output of a neuron or a layer, introducing non-linearity into the network. The standard activation functions utilized in CNNs are ReLU, sigmoid, and tanh.
- **Pooling**: It is a down sampling operation that reduces the spatial dimensions of the input feature map while retaining crucial information.
- **Fully Connected Layer**: A convention NN layer where each neuron is linked to each neuron in the preceding layer, capturing global information from the feature maps. It is typically used in the final stages of a CNN for classification, regression and segmentation tasks.
- **Feature Map**: The output of a convolutional layer, representing the learned features of the input image or previous layer.
- **Receptive Field**: The region of the feature map a neuron considers in a convolutional layer. It resolves the scope of spatial information that the neuron can capture from the data.

- **Dropout:** A regularization method to decrease the overfitting of the model. It establishes a random fraction of the input neurons to zero forcing the network to pick up substantial characteristics during training.
- **Batch Normalization**: It is applied to enhance the stability and training speed of CNNs. It normalizes the activations of each layer by standardizing them to have zero mean and unit variance.
- **Backpropagation:** It calculates the gradients of the network parameters concerning the loss function, allowing the network to learn through iterative optimization.
- **Loss Function:** It measures the distinction between the predicted output of the network and the actual output. The commonly used loss functions in CNNs include categorical cross-entropy, mean squared error (MSE), and binary cross-entropy.
- **Optimizer:** An algorithm that updates the network weights based on the gradients computed during backpropagation. The popular optimizers include Stochastic Gradient Descent (SGD), Adam, and RMSprop.

The flow chart used to segment the SR on the images is shown in Figure 5.3. The labelled images, i.e., binary masked images and the input images considered as the input in the CNN-based segmentation model, were used to predict the SR from the images.

5.3.1 Fully Convolution Neural Network for Segmentation

It is the extension of a *"Feed-forward neural network"* (FFNN) highly used for the semantic segmentation of digital images. It performs pixel-level predictions and end-to-end learning for segmentation tasks [127]. This model is playing a major role in the segmentation of the medical images due to its pixel wise segmentation properties [142][143]. The initial CNN model discards the spatial information, which makes the model unsuitable for pixel-level prediction. The fully linked layer of 1x1 convolutions is used in the FCN model to preserve the spatial data through the model training process to overcome the challenges. The FCN model is built with the encoder, decoder part, skip connection, fully convolutional output upsampling and loss functions:

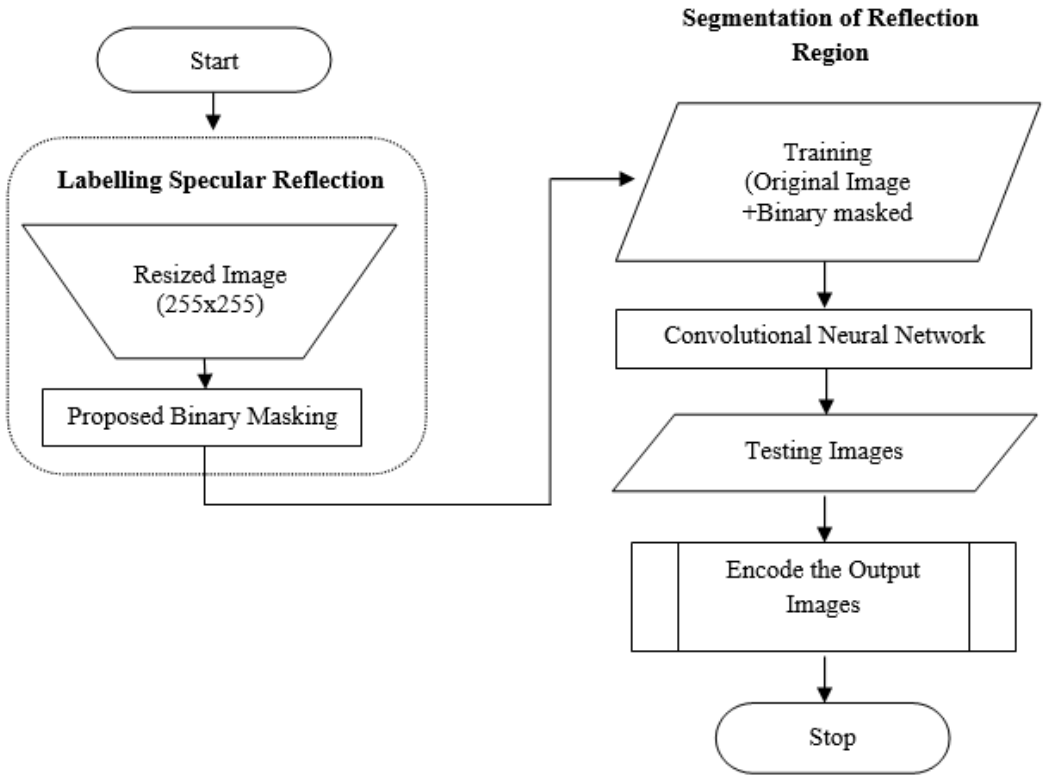


Figure. 5.3 Flowchart for the Segmentation of Specular Reflection using CNN Model

- **Encoder Network**: The FCN model is built with the pre-trained CNN model called VGG16, which is used in the encoder region. High-level features extracted from the input image by the encoder using pooling and convolutional layers. The feature captures the spatial information of the colposcopy images at different scales of the images.
- **Decoder Network**: The encoder is followed by the decoder network, designed to upsample the low-resolution feature maps obtained from the encoder to the original image resolution. This process is crucial for generating pixel-level predictions. The decoder network consists of transpose convolutions (deconvolutions) to perform upsampling.
- **Skip Connections**: The vital innovation introduced by FCN is the use of skip connections. It allows the model to utilize low-and high-level features during up sampling. This model creates the skip connection by combining the encoder network's feature maps with the decoder network's corresponding up sampled feature map. It helps preserve the fine-grained details in the up-sampling process to improve

the standard of the predicted output. It allows the decoder network to combine the low-level features acquired through the encoder, which have the high-level features where decoder learn to produce a more precise segmentation map.

- **Fully Convolutional Output**: Traditional CNN architectures that produce a single output prediction at the end, FCN generate dense pixel-wise predictions for semantic segmentation. The end layer of the FCN model is a 1x1 convolutional layer that extract a multi-channel output, where each channel indicates the probability of a particular class or the presence of a specific object at each pixel location.
- **Up sampling and Loss Function**: The up sampled output is usually lesser than the input image due to the encoder's pooling and stridden convolution operations. Bilinear interpolation or other up sampling techniques are used to obtain a prediction of the same size as the input image. It is typically trained utilizing BCE loss to compare the predicted segmentation map with the ground truth during training.

The first block of the FCN architecture typically includes of two convolutional layers, each with filters of 32. These filters are employed to the feature maps to acquire low-level features such as the digital images' edges, corners, and textures. The activation employed is ReLU, which presents non-linearity and helps learn complex patterns. The second block also consists of two convolutional layers, each with 64 aa filter size. In this layer, the model focuses on learning more sophisticated patterns and structures in the input image or feature maps. The third block includes two convolutional layers, each with 128 filters. These filters aim to capture even higher-level features and more complex spatial information. They are capable of identifying more abstract patterns and structures in the input image. The fourth block comprises two convolutional layers, each with 256 filters. The purpose of these filters is to capture even more advanced and abstract features. The fifth block consists of two convolutional layers, each with 512 filters. These filters are responsible for capturing the highest-level features and semantic information. They have a broader receptive field and aim to capture global details about objects and their relationships in the input. As an alternative of fully connected layers at the network's end, FCN employs 1x1 convolutional layers. These layers have a single filter per spatial location and are responsible for combining and aggregating information from different channels. It helps to preserve spatial details and allows the model to produce dense pixel-wise predictions. The model is built with the kernel size 3x3 with a stride value of 1. The dropout is set as 0.2, and

the batch normalization is enabled as true. The network architecture of the model is shown in Figure 5.4.

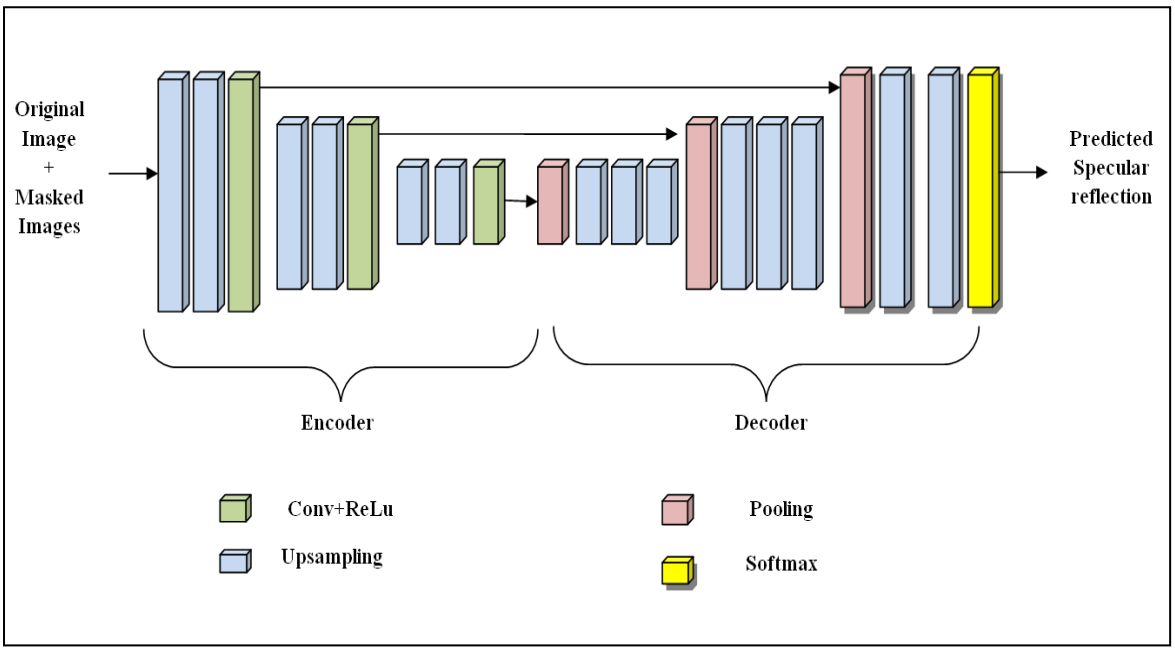


Figure 5.4 Network Architecture of the Fully Convolutional Network Model for the Specular Reflection Prediction

5.3.2 SegNet Model fo r Segmentation

SegNet is another popular architecture for semantic image segmentation [128]. It was proposed as an efficient and effective model for pixel-wise segmentation. An overview of the SegNet architecture for the segmentation of glare region from the images:

- **Encoder:** SegNet model starts with an encoder part that captures the image context. convolutional layers comprise the encoder, which is succeeded by ReLU activation and BN. The encoder layers advances the spatial dimensions by increasing the channels. It uses the max pooling layer with indices to store the locations of the maximum values, which are later used for up sampling in the decoder.
- **Decoder:** The decoder part in SegNet is accountable for upsampling the feature maps to the original image resolution. The series of transposed convolutional layers. It performs upsampling using the stored indices from the encoder's, associated with max pooling layers. This process allows SegNet to restore the spatial dimensions efficiently.
- **Fully Convolutional Output**: SegNet generates dense pixel-wise predictions for semantic segmentation. The final layer of the SegNet model is a 1x1 convolutional

layer that produces a multi-channel output. Each channel indicates the probability of a particular class or the presence of a specific object at each pixel location.

- **Upsampling and Loss Function**: It uses the stored indices from max pooling to upsample the feature maps during process. It helps in efficiently restoring the spatial dimensions of the images. For training, SegNet is typically trained using a loss function called binary cross-entropy due to the binary segmentation of the model.

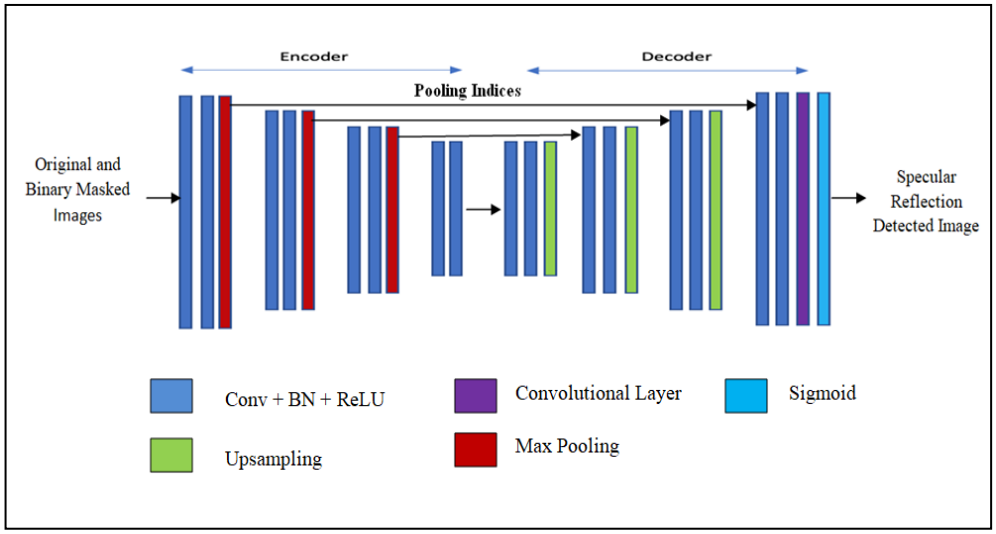


Figure 5.5 SegNet model for SR Segmentation

It uses the "*encoder-decoder*" structure, where each block in the encoder part gradually increases the number of filters while reducing the spatial dimensions. It is built with the five encoder blocks and four decoder blocks. The first block comprises of two convolutional layers, with a 64-filter size. The second encoder block also includes the two convolutional layers, each with a filter size of 128. Similarly, the third, fourth and fifth encoder blocks comprise four convolutional layers, each with a filter size of 256, 512, and 512, respectively. The decoder part of the SegNet model performs an upsampling operation to increase the spatial dimension while reducing the number of filters. The fifth and fourth block is built with the four-transpose convolutional with a filter size of 512. The third block comprises four transposed convolutional layers with filter size of 256. Similarly, the second and first blocks comprises of two transposed convolutional layers with a filter size of 128 and 64, respectively. This model does not use the skip connection; instead, it relies on the stored indices during the max pooling operation to guide the upsampling process. The 1x1 convolutional layers combine and aggregate each channel's information to preserve the spatial information during the segmentation process. The SegNet model for SR segmentation which is shown in Figure 5.5.

5.3.3. U-Net Model for Segmentation

It is generally used for biomedical image segmentation, where it has achieved higher efficiency in the segmentation method using smallest possible dataset [129][144]. Similar to the SegNet model, it comprises of an encoder path that acquire the image context and a decoder path that permits accurate location. Spatial information is preserved via the skip connections between corresponding levels in the encoder and decoder.

- **Encoder:** This architecture starts with an encoder part that consists of a series of pooling and convolutional layers. These layers are in charge of taking high-level features out of the input image. Typically, the encoder uses a series of downsample operations, such as max pooling, to increase the number of channels while decreasing the feature maps' spatial dimensions.
- **Decoder:** Following the encoder, the U-Net architecture employs a decoder part that consists of a series of transposed convolutional layers. These layers perform upsampling to restore the feature maps' spatial dimensions correspond to the original image resolution. The decoder gradually increases the spatial resolution while reducing the count of channels through transpose convolutions.
- **Skip Connections**: An essential aspect of this architecture is incorporating skip connections. It links the encoder and decoder layers at different resolutions. Low-level data is transferred from the encoder to the decoder via the skip connections, enabling the model to combine low-and high-level features during upsampling. It aids in preserving detailed spatial details and enhancing the segmentation accuracy. The model can gather both local and global contextual information to the encoder-decoder structure with skip links, producing precise and comprehensive segmentation maps.
- **Fully Convolutional Output:** Similar to FCN, U-Net generates dense pixel-wise predictions for semantic segmentation. A 1x1 convolutional layer, the last layer in the U-Net model, generates an output with several channels. Each channel represents the probability of a particular class or the presence of a specific object at each pixel location. The number of segmented objects is reflected in the total amount of channels in the output.
- **Upsampling and Loss Function**: During the upsampling process, the upsampled output from the decoder may be smaller than the original input image due to the

encoder's pooling and stride convolution operations. Bilinear interpolation or other upsampling techniques are commonly used to resize the output to match the input image size. For training, the U-Net model is typically trained using a loss function such as binary cross-entropy, which contrasts the segmentation map predicted by the model with the actual data. It is designed explicitly for semantic segmentation tasks and is known for its ability to handle limited training data.

It comprises of an encoder-decoder structure. Each block in the encoder part gradually increases the filters while reducing the spatial dimensions. The encoder block one usually comprises two convolutional layers with 64 filters. These filters capture lower features such as edges and textures in the input image or feature maps. The block 2 comprises of two convolutional layers, each with 128 filters. These filters focus on learning more complex patterns and structures. The block 3 includes two convolutional layers, each with 256 filters. The filters in this block aim to capture higher-level features and more abstract representations. The fourth block typically consists of two convolutional layers with 512 filters. These filters further extract advanced and abstract features. Usually, two convolutional layers are present in the fifth block, each with 1024 filters. These filters capture the highest-level features and semantic information from the smart colposcopy images. The decoder feature maps upsamples the original image resolution while reducing the number of filters. This block comprises two transposed convolutional layers, each with 512 filters. These transpose convolutions perform upsampling and help recover spatial information. The third block comprises of two transposed convolutional layers with 256 filters. These layers further upsample the feature maps while reducing the channels number. Block two usually cover two transposed convolutional layers, each with 128 filters. The filters in this block continue upsampling and reducing the channel numbers. The first block of the decoder typically comprises two transpose convolutional layers, each with 64 filters. These filters perform the final upsampling by reducing the channel numbers to match the desired output. Like FCN, U-Net uses 1x1 convolutional layers instead of fully connected ones. These layers have a single filter per spatial location and help combine and aggregate information from different channels. They aid in preserving spatial information and enable dense pixel-wise predictions. The UNet model for SR segmentation is shown in the Figure 5.6.

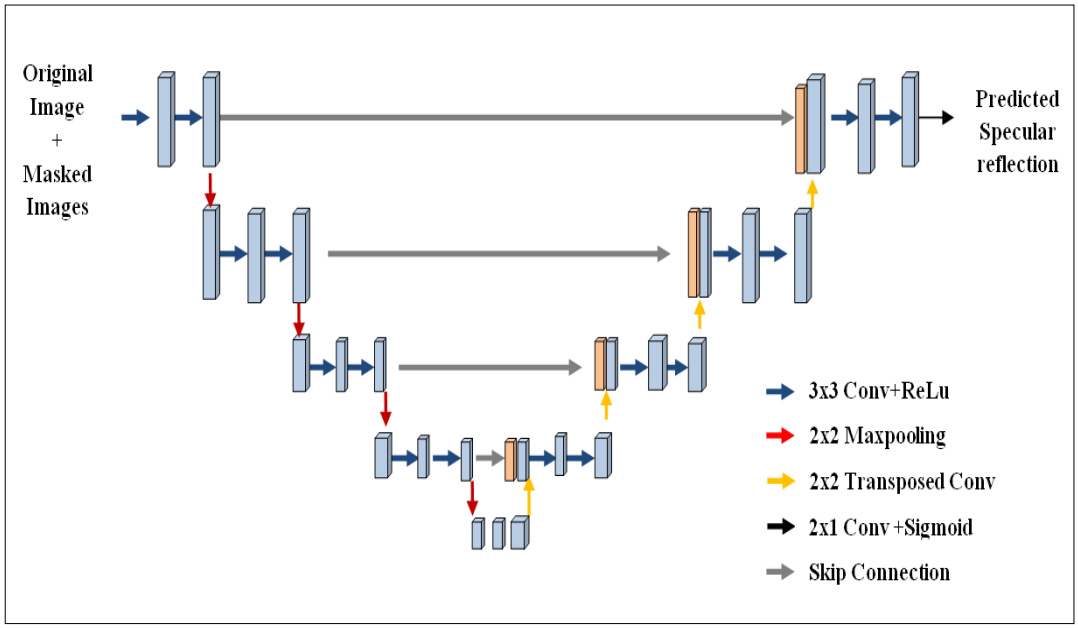


Figure. 5.6 UNet model for the SR Segmentation

5.4 Experimental Results and Discussion

The experimental setting required for the comparison analysis is addressed in this section. The *"Tesla V100 PCIe GPU and CUDA version 11.4"* were installed on the computer used to train the deep learning models, which provided the computational power necessary to train the models efficiently. The findings are conducted both quantitative and qualitative analysis for the segmentation models to predict SR on smart colposcopy images.

5.4.1 Training the Segmentation Model

To train the model for the SR segmentation on the images are:

- **Input images**: The images are resized to 255x255 pixels with three color channels (RGB).
- **Dataset sizes**: The training dataset has 3870 images, the testing dataset has 448 images, and the validation dataset has 448 images.
- **Model inputs**: The segmentation models take the original images and their respective masked images as inputs. These images are labelled to predict the SR from the smart colposcopy images.
- **Batch size and learning rate**: It is set as 32 for model training, which means that the optimizer updates the model's parameters after processing 32 images at a time. The learning rate is specified as 0.001, which controls the step size taken during parameter updates.

- **Optimizer:** The optimizer known as Adam is employed for training the model. It is known for its efficient computation process and have the capacity to tolerate parse gradients, making it suitable for medical images.
- **Loss function and activation function**: Binary cross-entropy is the employed loss function, and it compares the expected output with the ground truth labels. The sigmoid activation function, which converts the model's output to a probability range between 0 and 1, is employed in the model's last layer.
- **Training epochs**: The model is trained for 50 epochs, which means that the entire training dataset has to run through the model 50 times during training. The provided information outlines the necessary components and settings for training deep learning for specular reflection prediction in smart colposcopy images.

5.4.2 Qualitative Analysis

The deep learning segmentation models are compared for the qualitative analysis to identify the model suitable for the identification of SR on images. On qualitative analysis, the SR is detected more accurately utilizing the model U-Net than the other segmentation models. In Figure. 5.7 (a) represents the original images with SR. The specular reflection detected using the FCN model is illustrated in Figure 5.7. (b). On analysis, the detected SR predicts the non-SR region as reflection on the images. The SR detected using the SegNet model is despite in Figure 5.7 (c). On analysis, the SR with low intensity is not predicted in the SegNet model. The SR detected utilizing the U-Net model is shown in Figure 5.7 (d). On analysis, the SR pixels is accurately identified without influencing the other region of the images. The SR with low-intensity and high-intensity values are accurately identified in this model. The segmentation of SR utilizing the model is illustrated in Table 5.1.

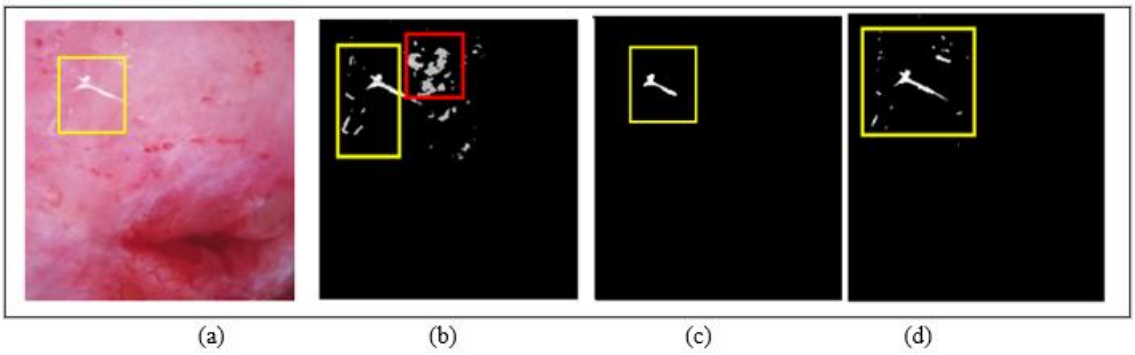


Figure 5.7 Comparison Analysis for Segmentation of SR. The *"red box"* indicates the non-SR predicted , and the *"yellow box"* indicates the SR on the images.

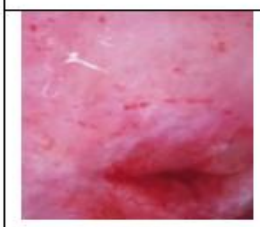
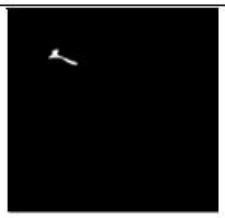
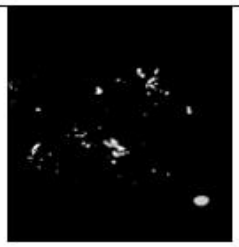
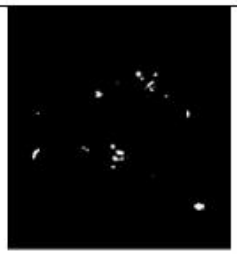
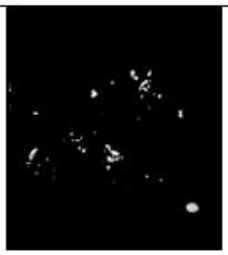
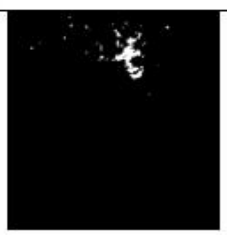
5.4.3 Quantitative Analysis

Binary accuracy is a commonly used metric to calculate the percentage of the correctly predicted pixel or region in the prediction binary mask in the segmented images [145].

$$BinaryAccuracy = \frac{Tp+TN}{TP+TN+FP+FN} * 100 \quad (5.2)$$

In equation 5.2, True Positives (TP) means the count of precisely predicted positive pixels. True Negatives (TN) represent the count of precisely predicted negative pixels. False Positives (FP) represent the count of wrongly predicted positive pixels. False Negatives (FN) represents the count of wrongly predicted negative pixels.

Table. 5.1 Segmentation of Specular Reflection on Smart Colposcopy using Convolutional Neural Networks

Original Image	Predicted Images Using Deep Learning Segmentation Model		
	FCN	SegNet	U-Net
			
			
			
			

Intersection over Union: It is referred to as the Jaccard Index, it is a well-liked assessment statistic for picture segmentation assignments [145]. It measures the overlap in between the predicted binary mask and the ground truth mask, as shown in equation (5.3)

$$IoU = \frac{Intersection}{Union} \tag{5.3}$$

Intersection indicates the count of pixels predicted as positive and present in the ground truth image. Union indicates the count of pixels predicted as positive or present in the ground truth mask.

Dice Coefficient: It is widely used metric for segmentation evaluation. It quantifies the agreement between the predicted binary mask and the ground truth mask, as shown in the equation (5.4).

$$DiceCoefficient = \frac{(2*Intersection)}{(Predicted+GroudTruth)} \tag{5.4}$$

Intersection indicates the count of pixels predicted as positive and present in the ground truth mask. In the equation 5.4, predicted indicates the total count of pixels predicted as positive. The ground truth represents the total number of pixels in the ground truth mask. It varies from 0 to 1, with 1 indicating an ideal overlap between the predicted and ground truth masks.

Loss Function: Loss function is used to scale the discrepancy between the predicted binary mask and the ground truth mask. In the case of binary image segmentation, BCE loss function is utilized.

$$BCELoss = -(Y * \log(Y_{hat}) + (1 - Y) * \log(1 - Y_{hat})) \tag{5.5}$$

In the equation 5.5, *y* represents the ground truth binary mask and *y_hat* represents the predicted probabilities of the binary mask, the *BCE loss* penalizes differences between the predicted and ground truth masks and encourages the model to learn accurate pixel-wise predictions. These metrics and loss functions are commonly used to evaluate and train binary image segmentation models to predict SR on the images. On quantitative investigation, the U-Net segments the SR with 98.73% of accuracy., as shown in Table 5.2

Table 5.2. Quantitative Evaluations for SR pixel Segmentation on Smart Colposcopy

Methods	Binary accuracy (%)	IoU	Dice Coefficient	Loss
SegNet	0.8246	0.8012	0.8417	0.2017
FCN	0.9781	0. 7984	0.8417	0.2017
U-Net	**0.9873**	**0. 7924**	**0.8679**	**0.1312**

The UNet model performs exceptionally well in predicting specular reflection, achieving a higher accuracy of 98.73%, as shown in Figure 5.8. The model's ability to accurately detect specular reflections is a testament to its effectiveness in analyzing smart colposcopy images. However, during the training process, the model faced specific challenges that led to moments of instability, resulting in a sudden fall in the training graph. Despite this temporary setback, the model recovered and exhibited a smooth and consistent improvement, ultimately delivering high accuracy. However, the sudden fall followed by a smooth line demonstrates the model's ability to adapt and overcome challenges, eventually converging to a high level of accuracy.

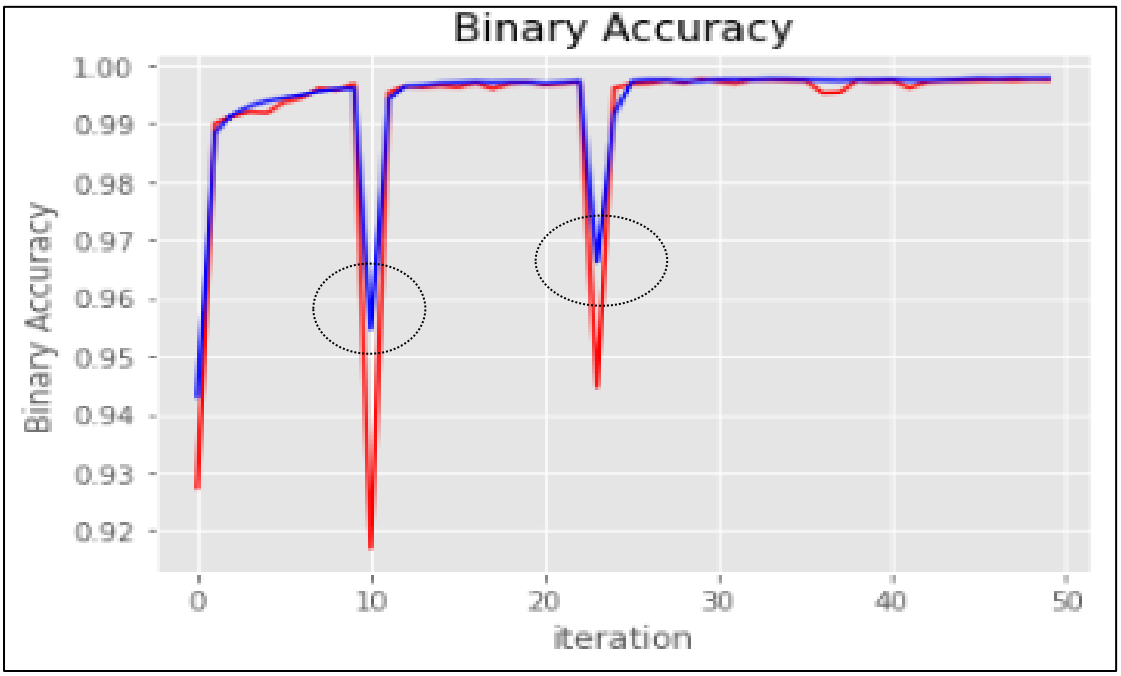


Figure 5.8. Binary accuracy for SR Pixel Segmentation on the Images using the UNet Model

The IoU for the segmented images should be higher because it shows the resemblance of the predicted and the ground truth images. An IoU value of 0 means no overlap in between the predicted and ground truth masks, indicating a complete mismatch in the images. On the other hand, an IoU value of 1 signifies a perfect match, where the predicted mask matches the ground truth mask. Even though the accuracy is 98.73%, the similarity of the predicted and the ground truth images is only 79.24%, as shown in the Figure. 5.9. It also shows slight overfitting of the model as represented in the Figure.5.9. The dice coefficient is often preferred for evaluating the similarity between predicted and ground truth segmented images because it provides a measure of overlap. A higher dice coefficient indicates a more significant similarity between the predicted and ground truth masks. Conversely, a dice coefficient value of 0 signifies a complete mismatch, indicating no mask overlap. On the other hand, if the value of the dice coefficient is 1, it represents a perfect match, where the predicted mask precisely matches the ground truth mask. The dice coefficient of the predicted specular reflection is 0.86, as shown in Figure 5.10, which represents 86% similarity to the ground truth images.

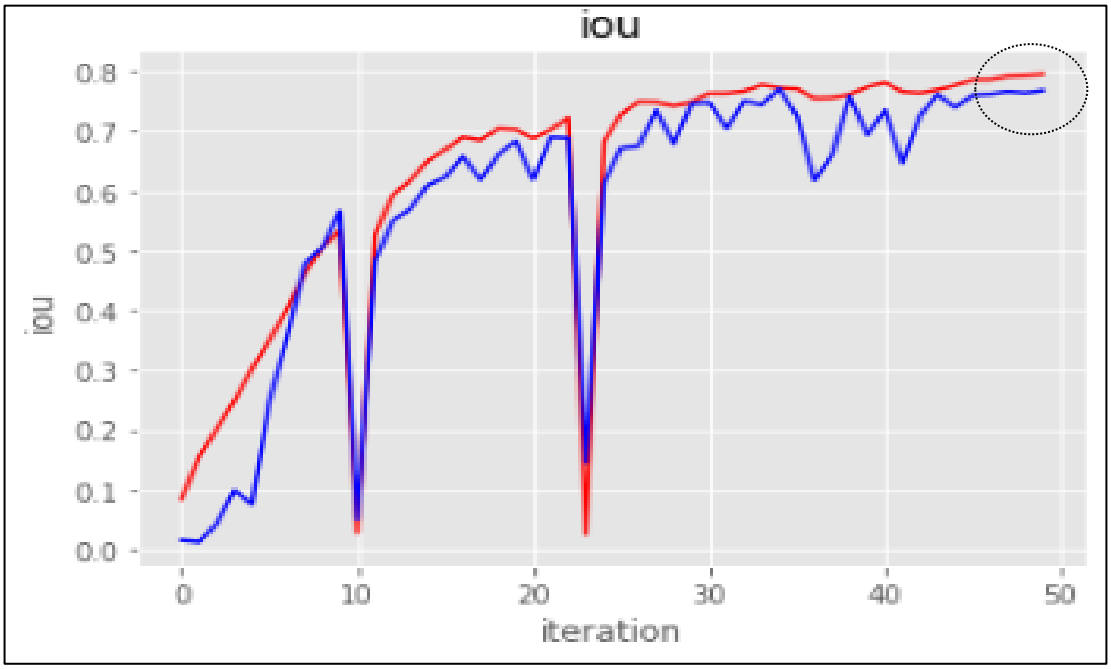


Figure 5.9. Intersection of Union for SR pixel Segmentation on the Images using the UNet Model

Figure 5.10. Dice Coefficient for SR pixels Segmentation on the Images

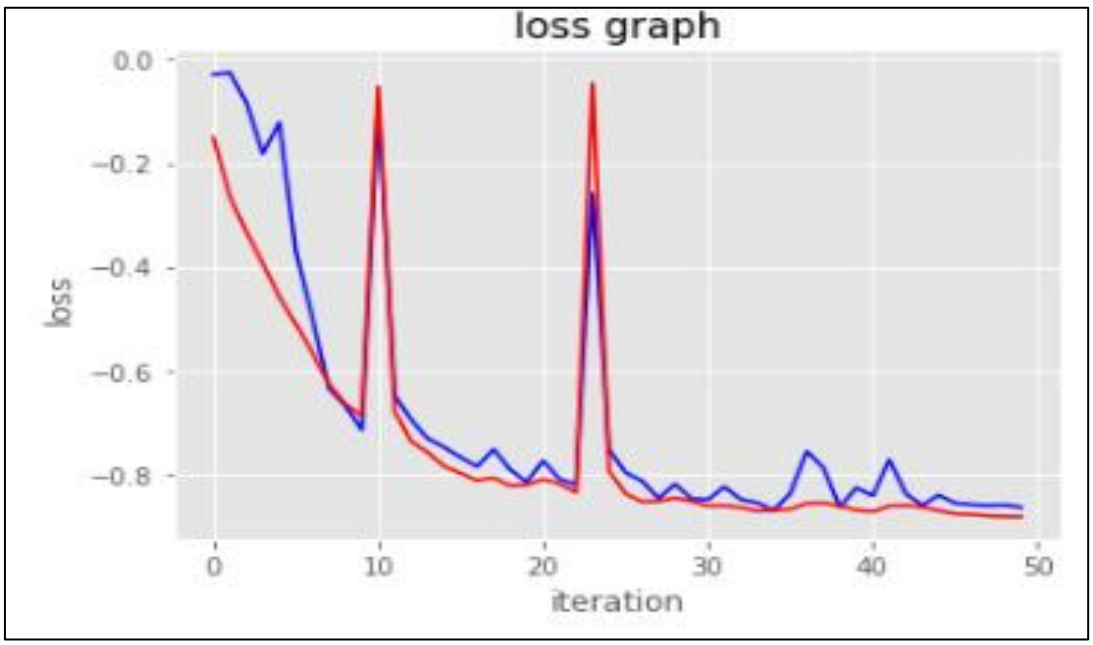


Figure 5.11. Loss Calculation for the SR Segmentation on Smart Colposcopy Images using UNet Model

The loss value is also calculated to predict the SR on the smart colposcopy images. The lower the loss value, i.e., 0, represents the clarity of the images, which is greater, and similarly, if the value is 1, the quality of the images is lower. The BCE loss is calculated for the predicted due to the binary pixel segmentation. On analysis, the SR detected on the images with a loss value of 0.1312, as shown in Figure 5.11. On analysis, the UNet model achieved an accuracy of 98.73% in predicting specular reflections using a dataset size of

3,870. However, despite the satisfactory accuracy, the IoU scored 79.24%, which did not meet the expected level. So, different versions of the UNet model were compared to identify the most suitable model for accurate prediction to enhance further the segmentation of SR in smart colposcopy images.

5.5. Segmentation of Specular Reflection on the Smart Colposcopy Images using Different Versions of UNet Model

The UNet model was examined, revealing that it accurately segments the SR with a remarkable accuracy of 98.23%. But there are different versions of the UNet model, like Residual UNet+ and Unet++ which are initially compared to predict the suitable model to predict the specular reflection with higher accuracy and similarity rate. So, different versions of the U-Net model are analyzed to improve the SR prediction on the images [130].

5.5.1 U-Net++ Segmentation Model for Specular Reflection

This architecture is an elaboration of the model U-Net that aims to maximize the accuracy and quality of image segmentation [82]. It incorporates the strengths of U-Net's encoding and decoding paths while introducing several novel components to enhance feature extraction and information flow throughout the network. It comprises of an encoding path, a decoding path, and a final set of convolutional layers. The encoding path gradually reduces the spatial dimensions of the input image, capturing hierarchical feature representations. The decoding path then upsamples the feature maps to reconstruct the segmentation mask with fine-grained details. Skip connections links the respective encoding and decoding layers to allow the direct flow of information across different resolutions.

Encoding Path:

- **Multi-resolution Fusion**: In U-Net++, the encoding path incorporates multi-resolution fusion, which enhances the feature depiction achieved by fusing information from different scales. At each encoding step, the feature maps from the previous layer are fused with the aligned upsampled feature maps from the decoding path. This fusion of multi-resolution features helps preserve contextual data that is both local and global.

- **Skip Connections**: Skip connections in U-Net++ facilitate the direct information flow from encoding to decoding layers. It mitigates the information loss due to downsampling in the encoding path.

- **Dense Skip Connections**: It addresses the dense skip connections to enhance the data flow. Unlike traditional skip connections that connect adjacent encoding and decoding layers, dense skip connections combine each encoding layer with all subsequent decoding layers. This dense connectivity pattern enables the sharing of information across multiple scales and promotes the formation of features from various tiers of abstraction.
- **Transition Layers:** Transition layers are employed in U-Net++ to regulate the count of feature channels during the transition from the encoding to the decoding path. These layers minimize the computational intricacy and prevent the network from overfitting by reducing the number of channels in the feature maps before upsampling.

Decoding Path:

- **Upsampling**: The decoding path in U-Net++ consists of upsampling operations, by increasing the feature maps' spatial resolution. Various upsampling techniques, such as transposed convolutions, can be employed to recover the lost spatial data during the encoding phase.
- **Concatenation:** To exploit the information from both the encoding and decoding paths. The upsampled data are combined with the corresponding feature maps from the encoding path using concatenation. This concatenated feature representation allows the network to leverage high-level semantics from the encoding layers while preserving the fine-grained details recovered during upsampling.
- **Skip Connections**: Similar to the encoding path, skip connections are utilized in the decoding path to enable the direct flow of information from previous decoding layers to next decoding layers. These connections help refine the segmentation mask by integrating features from multiple resolutions.
- **Transition Layers:** Transition layers in the decoding path are analogous to those in the encoding path. They regulate the count of channels in the feature maps before passing them through the final set of convolutional layers.
- **Final Convolutional Layers**: After the decoding path, U-Net++ incorporates a set of convolutional layers to generate the last segmentation mask. These layers typically consist of 1x1 convolutions after which each class's per-pixel probability ratings are generated using a sigmoid function. The UNet++ model for specular reflection is shown in Figure 5.12. The network architecture, filter and neuron size

are similar to the UNet Model for predicting glare region on the smart colposcopy images.

5.5.2. Residual UNet Model for Segmentation of Smart Colposcopy Images

The Residual U-Net architecture combines the U-Net model with residual connections inspired by the residual learning framework introduced in the ResNet architecture [146]. This integration aims to address the vanishing gradient problem and improve the flow of information during training, leading to enhanced feature representation and more accurate image segmentation. Residual U-Net combines skip connections within the encoding and decoding paths, enabling the direct flow of information across different layers. These connections allow for the propagation of gradients and facilitate the reuse of features from earlier layers, promoting better information flow and reducing the likelihood of information loss.

Encoding Path

- **Convolutional Layers:** The encoding path of Residual U-Net start with a number of convolutional layers that acquire features from the input image. These layers utilize filters of various sizes to capture different levels of contextual information.
- **Residual Blocks:** Residual blocks are the critical component of the encoding path in Residual U-Net. It consists of multiple convolutional layers succeeded by element-wise addition with the input to form a residual connection. This connection enables learning the residual features and facilitates the flow of gradients during training.

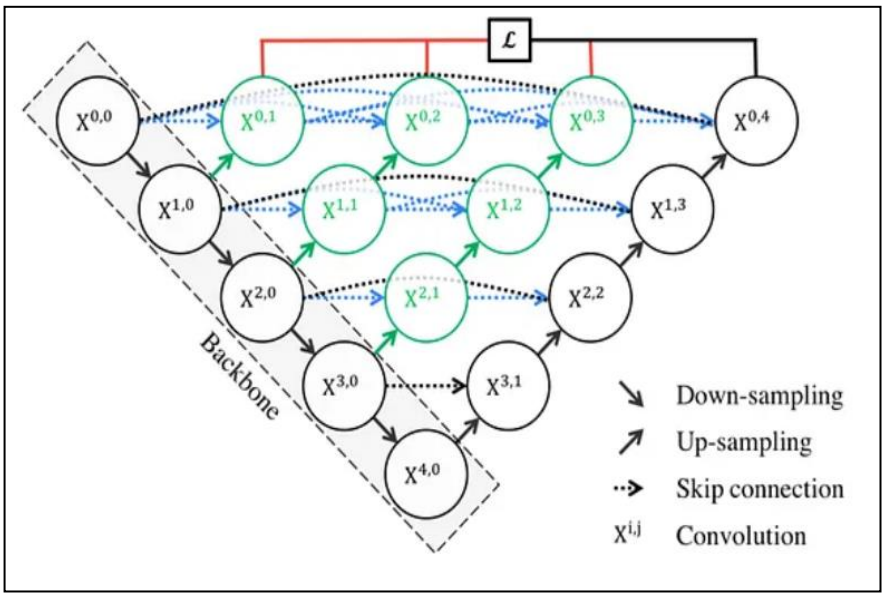


Figure 5.12 UNet++ Model for the Segmentation of SR .

Decoding Path

- **Upsampling**: The upsampling operations are used to restore the feature maps' spatial resolution. These techniques such as bilinear or transposed convolutions can be employed to increase the feature maps size.
- **Concatenation**: Similar to the original U-Net, Residual U-Net uses concatenation to fuse the upsampled data with the corresponding data from the encoding path. This concatenated feature representation ensures the integration of multi-scale information and allows the model to recover fine-grained details.
- **Residual Blocks**: Residual blocks are also employed in the decoding path to introduce the residual connections and facilitate the flow of information. Each residual block in the decoding path receives inputs from both the current decoding layer and the corresponding encoding layer. It enables the model to refine the segmentation mask by incorporating features from different resolutions.
- **Final Convolutional Layers**: Residual U-Net incorporates convolutional layers to generate the final segmentation mask. These layers typically consist of 1x1 convolutions followed by an activation function, called sigmoid, to produce per-pixel probability scores for each class.

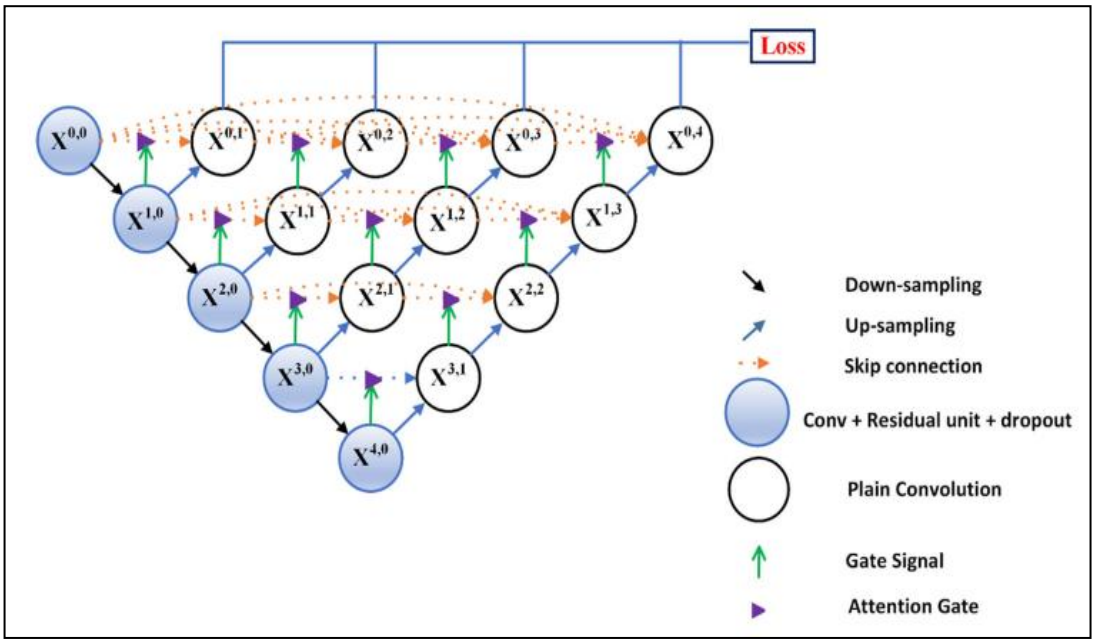


Figure 5.13 Residual UNet Model for the Segmentation of SR

- **Loss Function**: The choice of function in Residual U-Net depends on the specific segmentation task. The BCE is employed to measure the discrepancy between the segmentation mask and the ground truth, guiding the training process to enhance the model's capability. The network architecture, filter and neuron size are similar to the UNet model for predicting SR on the images. The residual UNet model for the SR segmentation is shown in Figure 5.13.

5.6. U-Net++ Fine-Tuned for SR Segmentation

The U-Net model has shown superior performance in segmentation tasks, particularly with minimal datasets in medical imaging [147]. In this study, the fine-tune the UNet++ model specifically for segmenting specular reflection in colposcopy images. The U-Net++ architecture, an extension of U-Net, is employed to enhance the model's feature representation and information flow. The U-Net++ model comprises an expanding path for feature decoding and a contracting path (encoder) for segmenting the regions of interest. The architecture consists of four blocks of convolutional, all containing two convolutional layers on the sides of encoder and decoder. The operation of convolutional layer can be represented by Equation 5.6.

$$F(i,j) = f\left(\sum_{m=0}^{2}\sum_{m=0}^{2} W_{m,n} I_{x+m,y+n} + W_b\right) (5.6)$$

Here, $W_{m,n}$ represents the weight of the m,n kernel, $I(x,y)$ represents the pixel values at position x,y, and $F(i,j)$ denotes the image feature map formed. Every block is succeeded by the ReLU function. Additionally, max pooling of $2x2$ succeeds all block in both the downsampling and upsampling paths. These certain modifications are made to fine-tune the U-Net++ model for glare detection in smart colposcopy images. Every single layer has batch normalization configured to hasten convergence when learning. A dropout value of 0.2 is applied to all layer in the region of down sampling. It has filter dimensions of 8, 16, 32, 64, and 128 in the contracting path and 128, 64, 32, 16, and 8 in the expanding path. The model is trained with a batch size of 10 to 50 epochs. The Adam optimizer is utilized with a learning rate of 0.00001. The BCE loss function is employed for two classes segmentation. By fine-tuning the U-Net++ model, which aim to improve reflection segmentation on smart colposcopy images, leveraging the enhanced feature representation and information flow facilitated by the UNet++ architecture. The algorithm to predict the SR using the fine-tuned UNet++ model is shown in the Algorithm. 5.2

5.7. Experimental Results and Discussion

The experimental setting required for the proposed model is discussed in this section. The deep learning models were trained on a machine equipped with a "*Tesla V100 PCIe GPU*" and "*CUDA version 11.4*", which provided the computational power necessary to train the models efficiently. The results are analyzed in the quantitative and qualitative analysis. The different versions of the UNet model are discussed in this section.

Algorithm 5.2 (Fine Tuned UNet++ Model)

Define *the UNet++ model with input shape (255, 255, 3)*
Initialize *the model with four convolutional blocks*

For each convolutional block:

> ***Add*** *a 3x3 convolutional layer with the ReLu activation function*
> ***Add*** *a 3x3 convolutional layer with the ReLu activation function*
> ***Add*** *a 2x2 max pooling layer*
> ***Add*** *dropout with a value of 0.2*
> ***Add*** *an expanding path to decode the features of the digital images*

For each expanding block:

> ***Add*** *a 2x2 up-sampling layer*
> ***Add*** *a 3x3 convolutional layer with the ReLu activation function*
> ***Add*** *a 3x3 convolutional layer with the ReLu activation function*
> ***Add*** *dense connections to connect with all preceding expanding blocks*
> ***Add*** *nested skip connections to connect with corresponding encoding blocks*
> ***Add*** *an output layer with a sigmoid activation function*

Compile *the model with binary cross-entropy loss function and Adam optimizer with a learning rate of 0.00001*

Set *the batch size to 10 and the number of epochs to 100. Set the input image size to (255, 255, 3)*
Load *the training and testing datasets with sizes (3870, 255, 255, 3 (Original and Masked Images)) and (1000, 255, 255, 3 (Original and Masked Images)) respectively*

5.7.1. Training the Segmentation Model

- **Input images**: The input images are resized to 255x255 pixels with three color channels (RGB).

- **Dataset sizes**: The training dataset has 3582 images, the testing dataset has 448 images, and the validation dataset has 448 images.

- **Model inputs**: The U-Net and its variation model take both the original images and their corresponding masked images as inputs. The masked images are labelled to predict the SR from the images
- **Batch size and learning rate**: The model is learned with 32 as batch size, which means that the optimizer updates the model's parameters after processing 32 images at a time. The 0.001 is set as the learning rate, which controls the step size taken during parameter updates.
- **Optimizer:** The Adam optimizer is employed for model training. It is known for its efficient computation process and ability to handle noisy or sparse gradients, making it suitable for medical images.
- **Loss function and activation function**: The loss function employed is binary cross-entropy, which compares the predicted output with the ground truth labels. The activation function employed in the last layer of the model is sigmoid, which convert the model's output into the probability interval between 0 and 1.
- **Training epochs**: The model is trained for 50 epochs, which means that the entire training dataset has to go through the model 50 times during training.

5.7.2. Qualitative Analysis

The different versions of UNet models are compared to identify the model suitable for SR segmentation. In qualitative analysis, the visual inspection method predicts the suitable model for SR segmentation on the images.

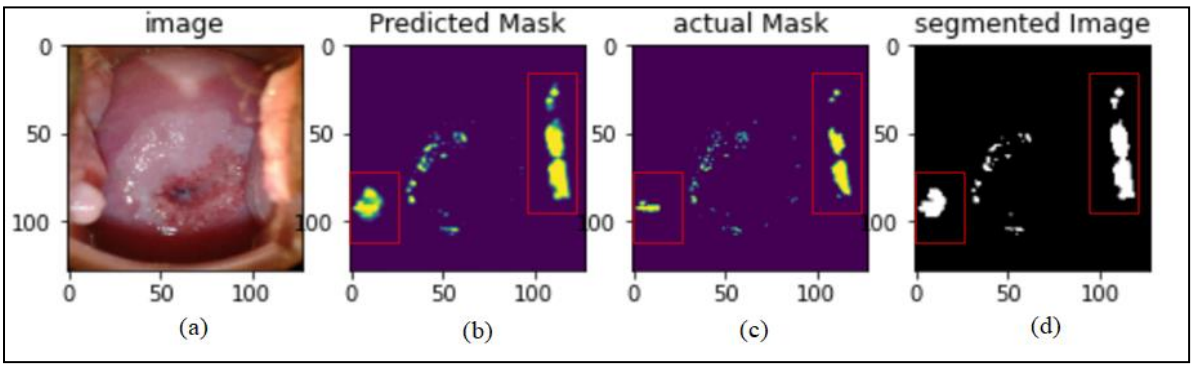


Figure5.14 Specular Reflection Segmented using the UNet Model. (a) Original Images (b) Predicted Mask (c) Actual Mask (d) Segmented Images

The UNet predict the specular reflection but, based on the analysis of the actual and predicted masks. The red-marked region shows that some of the non-specular reflections are

identified as specular reflections in the predicted mask as shown in the Figure 5.14(c). The segmented image also indicates that non-glare regions are segmented from the images, as shown in Figure 5.14(d).

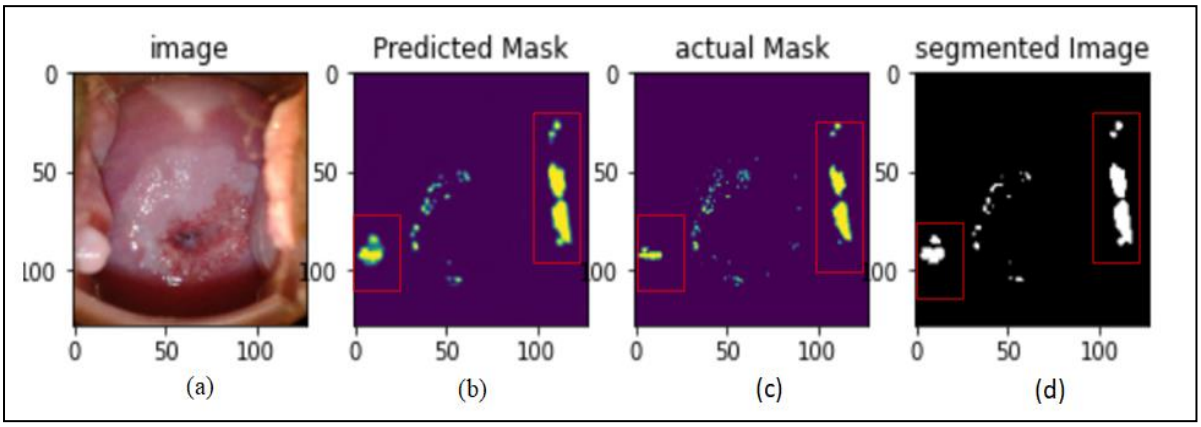


Figure 5.15 Specular Reflection Segmented using the UNet++ model. (a)Original Images (b) Predicted Mask (c) Actual Mask (d) Segmented Images

The UNet++ predicts the SR, but based on the analysis of both the actual and predicted masks. The analysis indicates that the UNet++ model outperforms the UNet model in this task, but still, some of the border regions are over segmented reducing the segmentation accuracy of the model as shown in the Figure.5.15(d)

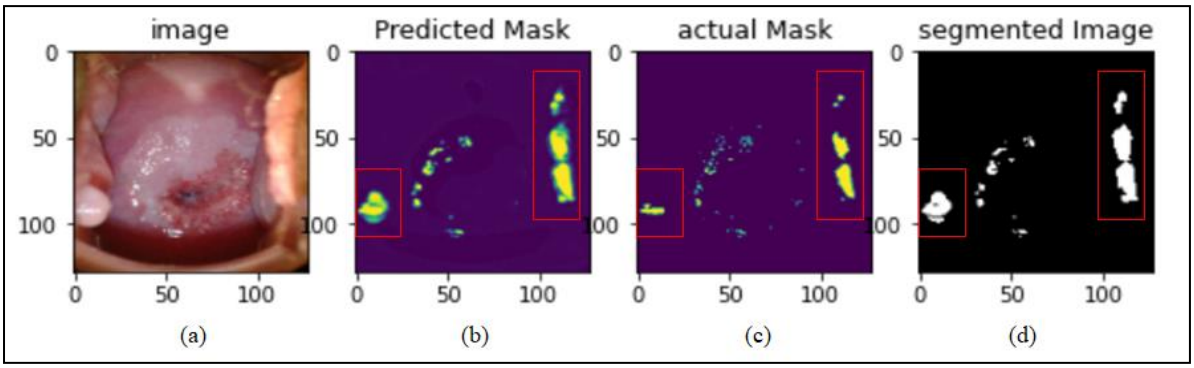


Figure 5.16 Specular reflection segmented using the Residual UNet Model. (a)Original Images (b) Predicted Mask (c) Actual Mask (d) Segmented Images

The Residual UNet over segments the certain area on the images as shown in Figure.5.16 (b). The segmented images show the non-SR region appears in the border of the specular reflection region is also segmented in the images using this model, as shown in Figures 5.16 (d). The Fine-tuned UNet++ model segments certain area on the images, as illustrated in Figure 5.17 (b). On visual analysis, the UNet ++ fine-tuned segments the SR without affecting the border of the reflection region as shown in Figure 5.17 (d). So, based

on the analysis, this model outperforms the other versions of the UNet model in identifying reflection regions on the smart colposcopy images.

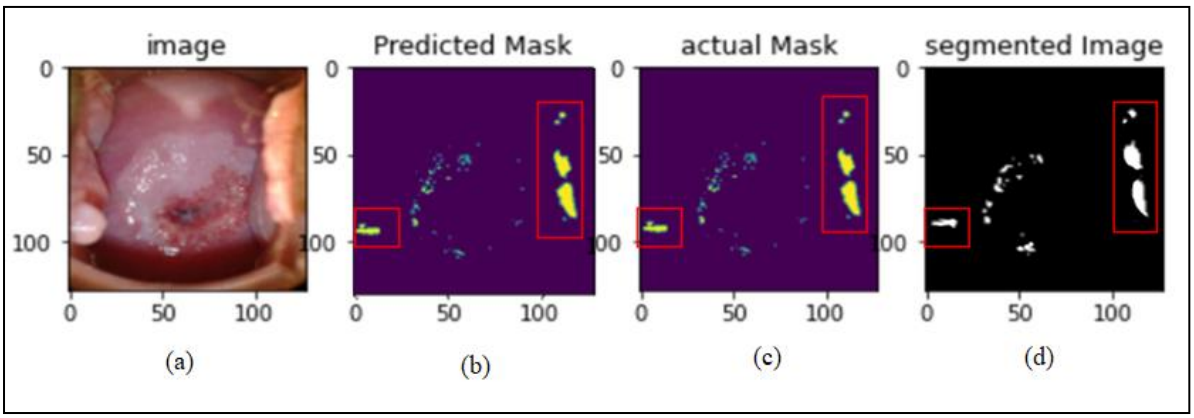


Figure 5.17 Specular Reflection Segment using fine-tuned UNet++ model. (a)Original Images (b) Predicted Mask (c) Actual Mask (d) Segmented Images

5.7.3. Quantitative Analysis

In quantitative analysis, these metrics and loss values are commonly utilized to assess the performance of segmentation models. Higher the values for binary accuracy, IoU, and the dice coefficient indicate better segmentation accuracy, while lower loss values indicate a closer match between the predicted and ground truth masks. These metrics and loss values provide objective measures of the model's performance and can be used to compare different models or track the model's progress during the training and validation phases. The quantitative analysis of the different versions of UNet models is compared to predict the suitable model for specular reflection segmentation on smart colposcopy images. The UNet++ model improved segments the SR with a greater accuracy of 99.98% on comparison analysis, as shown in Table 5.3.

Table 5.3 Quantitative Evaluation for the Segmentation of SR on the Images

Original	Binary Accuracy	Intersection of Union	Dice coefficient	Loss
UNet	98.73	0.7924	0.8679	0.1312
UNet++	99.23	0.9328	0.9012	0.0132
Residual UNet+	81.36	0.6212	0.7914	0.1562
Fine Tuned UNet ++	**99.98**	**0.9977**	**0.9341**	**0.0082**

The Fine-tuned UNet++ model demonstrates exceptional performance in predicting specular reflection, achieving a higher accuracy of 99.98%, as shown in Table. 5.3. The model's ability to accurately segment specular reflections without overfitting or spike fall in the graph. The binary accuracy reached a higher accuracy with the epoch of 20 as shown in the Figure 5.18. The dice coefficient is commonly used to assess the similarity between predicted and ground truth segmented images, as it provides a measure of overlap. A higher dice coefficient indicates a more remarkable resemblance between the predicted and ground truth masks. A dice coefficient value of 0 signifies a complete mismatch, indicating no overlap between the masks, while a value of 1 represents a perfect match, where the predicted mask precisely aligns with the ground truth mask.

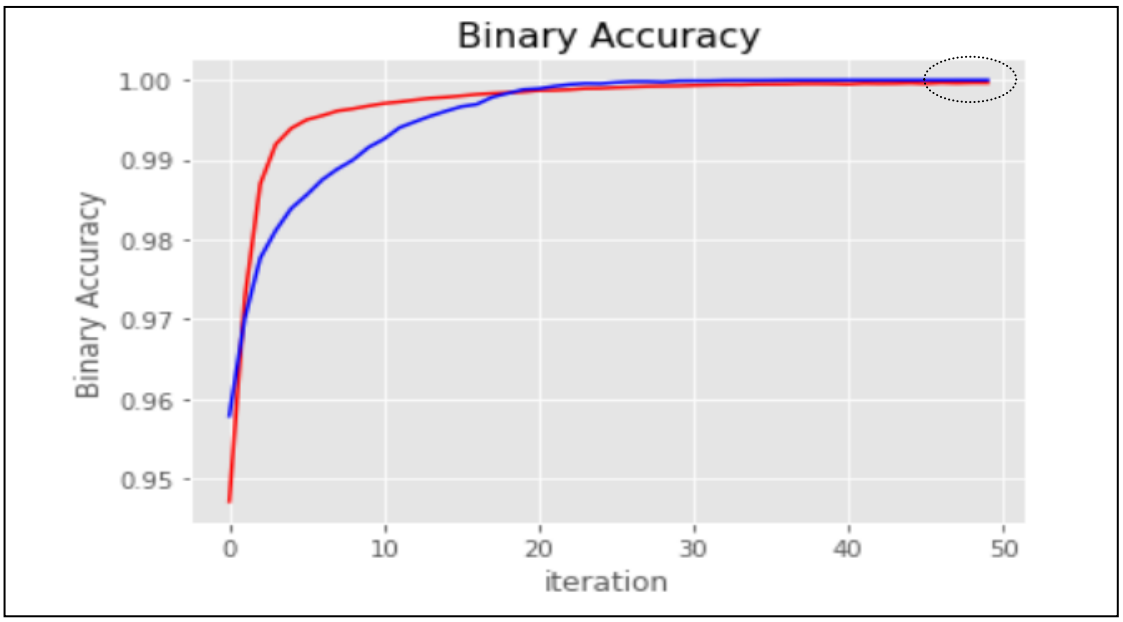


Figure 5.18 Binary Accuracy for the Prediction of Specular Reflection on Smart Colposcopy Images using Fine-tuned UNet++ Model.

In comparison analysis of the fine-tuned UNet++, the dice coefficient for the predicted SR is 0.93, indicating a substantial similarity between the predicted and ground truth masks, as shown in Figure 5.19. Similarly, as shown in Figure 5.20, the segmentation of SR resulted in a loss value of 0.0082. It measures the overlap between these two masks and provides insights into their similarity. A higher IoU value indicates a more significant overlap and similarity between the predicted and ground truth masks, with a value of 1 representing a perfect match. Conversely, an IoU value of 0 indicates no overlap, indicating a complete mismatch between the masks.

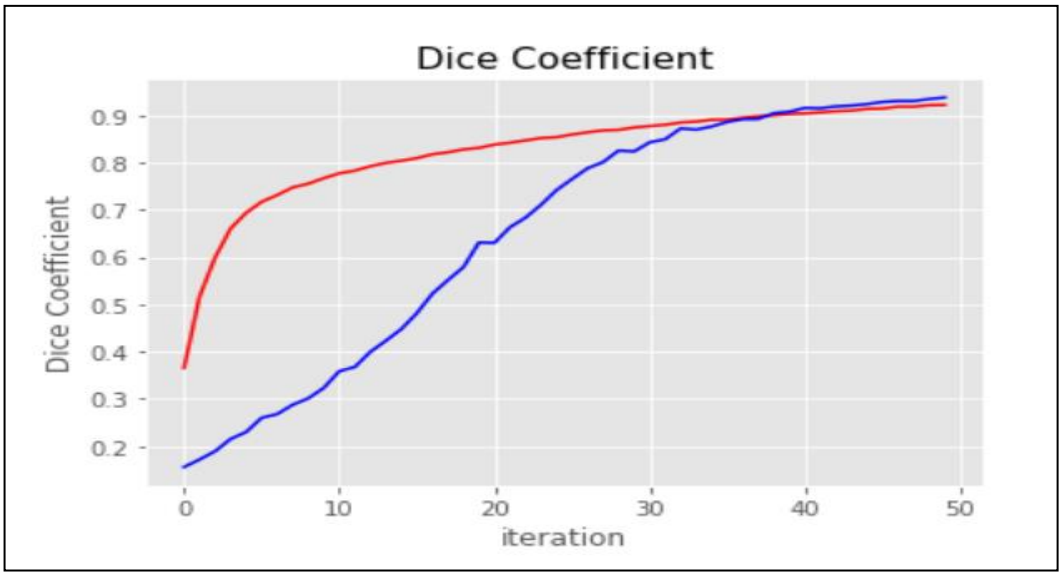


Figure 5.19 Dice Coefficient for SR Segmentation using Fine-tunedUNet++ Model

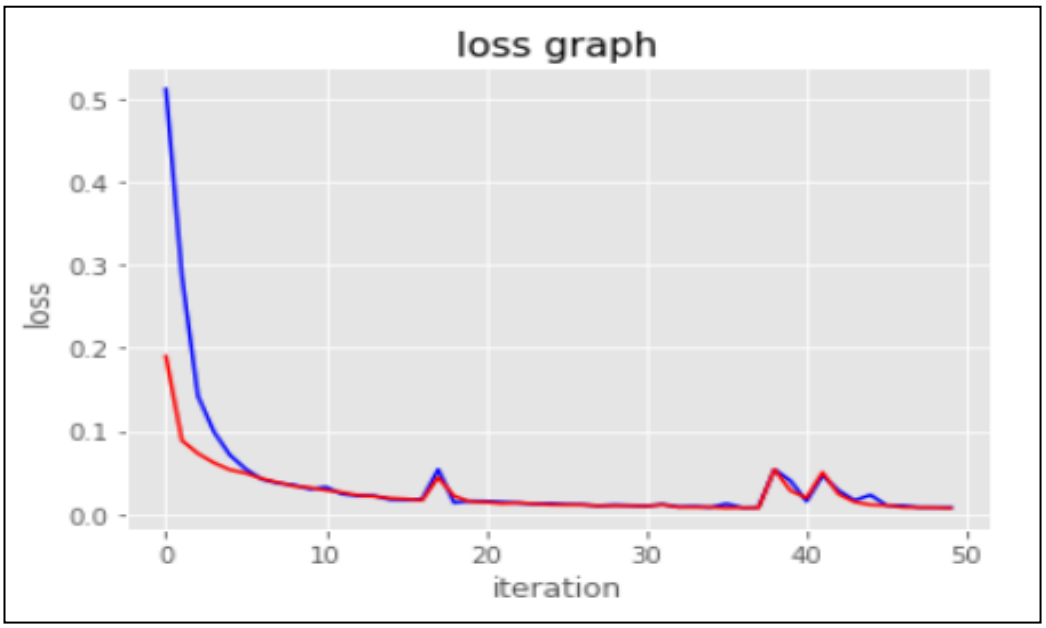


Figure. 5.20 Loss Graph for SR Segmentation using Fine-tuned UNet++ Model

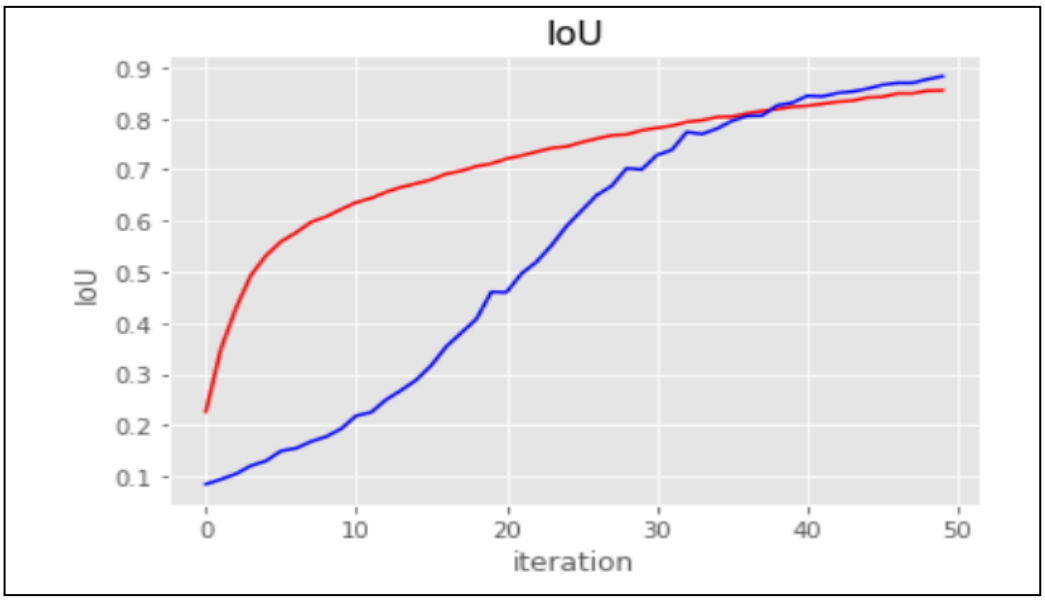


Figure 5.21 IoU for SR Segmentation using Fine-tuned UNet++ Model.

The IoU predicts the of SR region with 0.9977, as shown in Figure 5.21. It shows that the predicted images using the fine-tuned UNet++ model and the actual images have a similarity of 99.77%.

5.8. Summary

The proposed approach initiates the generation of binary mask images through a threshold method, which effectively creates annotated images that serve as the basis for segmenting images using deep learning techniques. This aspect is crucial as it enables the model to identify and differentiate between different regions of interest within the images. A comparative analysis is performed among three-pixel wise segmentation models, SegNet, FCN, and UNet, to determine the most suitable deep learning segmentation model. The evaluation focuses on their performance in accurately segmenting the images and achieving the desired results. Among these models, UNet emerges as the top performer, showcasing superior capabilities in segmenting specular reflection on the smart colposcopy images. However, UNet demonstrates promising results; the achieved IoU metric must meet the expectations for accurate segmentation. To address this, further analysis is conducted by exploring different variations of the UNet model. The UNet++, and Residual UNet are used for the analysis for the segmentation of images. During the analysis, UNet++ exhibited exceptional prediction accuracy and IoU metric performance. The metrics indicate the model's ability to precisely outline and distinguish the glare regions within the smart colposcopy images.

Consequently, UNet++ is the preferred choice for further refinement and improvement. A fine-tuning process enhances the UNet++ model's suitability for glare segmentation in smart colposcopy images. This fine-tuning stage involves adjusting and optimizing various parameters, hyperparameters, and network architectures specific to the UNet++ model. The objective is to fine-tune the model to attain even higher accuracy, robustness, and efficiency in glare segmentation. Following the fine-tuning process, the UNet++ model demonstrates notable accuracy and IoU metrics advancements when predicting glare regions within the smart colposcopy images. This significant enhancement confirms the model's improved suitability for glare segmentation, offering more reliable and precise identification of these critical areas. The proposed method employs a threshold-based approach for generating annotated images, enabling deep learning-based segmentation. Through a comparative analysis, the UNet++ model proves its superiority over other segmentation models, surpassing expectations in accuracy and performance. After fine-tuning, the UNet++ model further enhances its accuracy and IoU metrics, explicitly excelling in segmenting glare regions within smart colposcopy images.