1.5em 0pt

A Comprehensive Study on Traditional as well as Deep Learning based Image Segmentation Methodologies

Rudranil Sardar (001910801070), Abhijit Das (001910801074),
Nayan Saren (001910801078), Rajarshi Bhattacharya (001910801080)
B.E.E. IV, Department of Electrical Engineering, Jadavpur University, Kolkata, India

***Abstract*—In this project, we first survey several semantic segmentation methodologies that can be employed for the task of binary segmentation. We choose a few thresholding based algorithms and deep learning algorithms and experiment them on the ACDC dataset in order to compare their efficacies. Visual comparisons as well as quantitative evaluations have been provided in order to make the survey more comprehensive and exhaustive. The code implementation for the project can be found at: https://github.com/rB080/fgd_Seg.git**

***Index Terms*—Artificial Neural networks, CNN, Thresholding, Image Segmentation, Deep Learning**

I. Introduction

Over the years, Deep Convolutional Neural Network (DCNN) architectures have significantly evolved for visual recognition tasks. This progress can be attributed to their end-to-end trainability. The breakthrough by Krizhevsky et al. [1] in supervised training of a deep model was a significant turning point for DCNNs. Despite being previously underutilized due to the extensive training data and computational cost required, models such as VGG [2], ResNet [3], Inception [4], MobileNets [5], and EfficientNets [6] have since emerged, achieving remarkable success in image classification, particularly on the ImageNet dataset.

CNNs have been primarily used for image classification since their inception. The network outputs the label that corresponds to the class of the object in the image. However, in certain visual tasks, such as biomedical image processing, precise detection and localization are crucial. In such cases, pixel-wise classification or image segmentation becomes essential. Semantic segmentation assigns a class label to each pixel in the input image, allowing for accurate localization instead of merely abstracting spatial details. Semantic segmentation has a vast range of applications, including scene understanding, autonomous driving, and biomedical image processing.

Several architectures have been proposed for the task of semantic segmentation, including FCN, U-Net [7], and UNet++ [8]. FCN [9] replaces the fully connected layers of a FCN classifier with convolution kernels and upsampling operations, allowing it to output a spatial map with pixel-wise prediction, which results in an efficient end-to-end trainable network. The architecture of FCN also incorporates skip connections between the downsampling and upsampling layers to combine deep, coarse, semantic information with shallow, finer features. U-Net, proposed by Ronneberger et al [7]. as a modification of FCN architecture, works well with fewer training samples and yields more precise segmentation. The large number of feature channels in the upsampling network of U-Net enables the network to propagate context information to higher resolution levels, resulting in symmetric downsampling and upsampling halves.

In UNet++ [8], the long skip connections are nested and densely connected to improve the U-Net architecture for image segmentation tasks [10]. This modification enriches the high resolution encoder feature maps before their aggregation with the semantically rich feature maps of the decoder, making the maps semantically similar. The nested and densely connected skip connections in UNet++ improve the propagation of multi-scale contextual information to the decoder, resulting in better performance in image segmentation tasks.

One approach proposed by Causey et al. [10] for kidney tumor segmentation is the use of an ensemble of two U-Net architectures, with one model predicting the kidney and tumor masks separately, and the other predicting the combined kidney-tumor mask along with the tumor mask. Meanwhile, a multi-scale feature aggregation network was reported by [11], which incorporates the dilated-inception block, consisting of four dilated convolutional layers. The outputs of these layers are concatenated and passed through a bottleneck layer for feature aggregation.

Cell segmentation is a common task in biomedical image analysis, and many methods proposed in the literature employ a single-stream segmentation pipeline. Despite the use of residual connections to counter problems of vanishing and exploding gradients, there remains a challenge to generalize the effective countermeasures against these issues.

For the task of optic disc (OD) and optic cup (OC) segmentation, Tabassum et al. [12] proposed an encoder-decoder architecture with a densely connected encoder on a SegNet backbone. However, these types of architectures often encounter a problem where feature resolution is reduced by consecutive convolution operations followed by pooling, impacting the learning of abstract feature representations.

Liu et al. [13] proposed an atrous CNN framework with a pyramid filtering module to obtain multiscale spatial aware features for the segmentation of the OC, which is more challenging than that of the OD due to its spatially sparse boundaries. Their model was tested on the DRISHTI-GS

dataset. In a separate study, Cao et al. [14] used the transformer architecture to improve the accuracy of 2D medical image segmentation, inspired by Swin Transformers. Their approach involves an encoder, bottleneck, decoder, and skip connections. Non-overlapping patches of input images are fed into the encoder to extract deep feature representations. The decoder up-samples the context features with a patch expanding layer and fuses them with multi-scale features from the encoder via skip connections to restore the feature maps' spatial resolution. Although this method reduces the inductive bias of CNN-based methods, it results in a loss of explainability.

In their study, Chen et al. [15] incorporated the self-attention mechanism for medical image segmentation as a sequence-to-sequence prediction task. To address the feature resolution loss brought about by Transformers, they proposed a hybrid CNN-Transformer architecture that utilizes both detailed high-resolution spatial information from CNN features and the global context encoded by Transformers. The self-attentive features encoded by Transformers are upsampled and combined with different high-resolution CNN features skipped from the encoding path, resulting in precise localization. This approach decreases the inherent inductive bias by employing a transformer-based encoder. Another study, [16], introduced an architecture that combines interleaved convolution and self-attention. It includes a lightweight convolutional embedding layer preceding transformer blocks.

The Unet architecture was improved upon by the authors in [17], who introduced the MultiRes block to replace the double convolution operation, providing a way to control the number of parameters. Res paths consisting of convolutional blocks replaced skip connections. A frequency level attention mechanism was introduced in [18], which uses a weighted combination of different frequency information types to aggregate the representation space. In [19], a network with advanced attention mechanism and multi-scale feature extraction modules was proposed to extract and aggregate multi-scale information at the feature level. A fusion model with an ensemble learning scheme was used to capture multi-scale information at the image level.

II. Dataset Description

We use the ACDC 2017 dataset [20] for our study. The dataset consists of volumes of Cardiac MRI Images from 100 patients and has a total of 951 slices (2D Images). From this, we use 900 slices for training set and 51 slices for test-set. All results reported in this study are on the test set, while the trianing set is strictly used for training the deep learning networks that we have studied.

III. Methodology

A. Deep Learning Methods

1) Unet: Unet [7] is a convolutional neural network (CNN) architecture that was proposed for biomedical image segmentation. The name "Unet" comes from the shape of the network, which resembles the letter "U." The architecture has been widely adopted and adapted for various segmentation tasks, such as in medical imaging, satellite imagery, and more.

The Unet architecture consists of two main parts: the encoder and the decoder. The encoder is a typical CNN architecture that is used to extract features from the input image. The decoder is used to generate a segmentation map from the extracted features. The encoder and decoder are connected by skip connections that pass information from the encoder to the decoder. This allows the decoder to use information from different scales to generate the segmentation map.

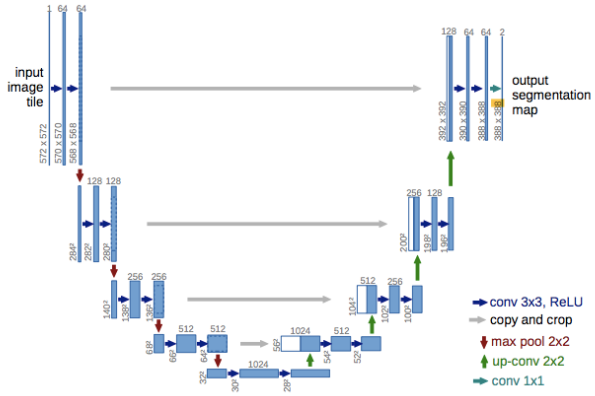


Fig. 1: The UNet architecture.

The encoder is made up of several convolutional layers, each followed by a rectified linear unit (ReLU) activation function and a max-pooling operation. The number of filters in each layer is doubled at each step, which allows the network to extract more complex and abstract features at each layer. The output of the final convolutional layer in the encoder is a set of feature maps that represent different scales and levels of abstraction.

The decoder is made up of several upsampling and convolutional layers. The upsampling layers are used to increase the resolution of the feature maps, while the convolutional layers are used to combine the high-resolution feature maps with the lower-resolution feature maps from the encoder. This is achieved using skip connections that concatenate the feature maps from the encoder with the corresponding feature maps from the decoder.

The final output of the Unet is a segmentation map, which is a pixel-wise classification of the input image. Each pixel in the segmentation map is assigned a label indicating the class to which it belongs. During training, the network learns to predict the correct label for each pixel by minimizing a loss function, such as binary cross-entropy or dice loss.

Overall, Unet is a powerful and flexible architecture that can be adapted to various segmentation tasks by modifying the number of layers, the number of filters, or the loss function used for training. Its effectiveness has been demonstrated in numerous biomedical image segmentation tasks, such as segmenting cells, organs, and tumors from medical images.

2) UNet++: Unet++ [8] is an extension of the popular Unet architecture for image segmentation. It was proposed by Zhou et al. to improve the segmentation accuracy of Unet. Unet++ builds upon the idea of skip connections in Unet and enhances it by adding nested skip pathways to improve feature

representation. The components of UNet++ can be explained in several steps:

1) Encoding: The UNet++ architecture starts with an encoding step, where the input image is processed through a series of convolutional layers to extract the features at multiple scales. The encoder network captures the features of the input image and reduces the spatial resolution of the feature maps.
2) Decoding: In the decoding step, the decoder network uses the features captured by the encoder to produce a segmentation map. The decoder network consists of a series of deconvolutional layers that upsample the feature maps and recover the spatial resolution lost during the encoding step.
3) Nested Structure: The UNet++ architecture incorporates a nested structure of multiple levels of U-shaped subnetworks, each of which contains an encoder and decoder network. The nested structure allows for the incorporation of skip connections between multiple layers in each subnetwork, enabling the network to capture multi-scale contextual information and enhance the feature extraction process.
4) Skip Connections: The skip connections in UNet++ allow for the flow of information between the encoder and decoder networks at multiple scales, which enables the network to better capture fine details in the input image. The skip connections also prevent the loss of spatial information during the encoding and decoding steps.
5) Weighted Aggregation Module: The UNet++ architecture incorporates a weighted aggregation module that enhances the fusion of features from different scales. This module allows the network to better balance the contribution of features from different scales and to effectively combine low-level and high-level features to produce more accurate segmentation maps.
6) Deep Supervision: The UNet++ architecture also incorporates deep supervision, which involves adding auxiliary segmentation outputs at intermediate layers of the decoder network. This helps to ensure that the network learns to produce accurate segmentation maps at all levels of the decoder network, further improving the accuracy of the segmentation results.

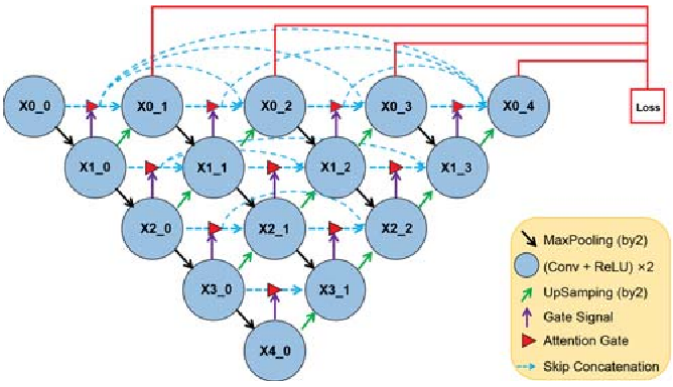


Fig. 2: The UNet++ architecture.

The UNet++ architecture's ability to capture multi-scale contextual information, balance the contribution of features from different scales, and incorporate deep supervision has made it a popular choice in the research community for image segmentation tasks. The architecture has been shown to outperform the original UNet architecture and other state-of-the-art segmentation networks in various image segmentation tasks.

3) FPN: The Feature Pyramid Network (FPN) [21] is a highly popular deep learning architecture that has found widespread application in computer vision, particularly in object detection and semantic segmentation tasks, as illustrated in Figure 3.

The fundamental idea behind FPN is to leverage the strengths of both two-stage and one-stage object detectors while mitigating their limitations. In two-stage detectors like R-CNN, high-level features from the backbone network are used to detect objects, whereas in one-stage detectors like YOLO, features from multiple scales are employed for detection. FPN combines these strengths by utilizing a pyramid of features from different scales to enhance the accuracy of object detection.

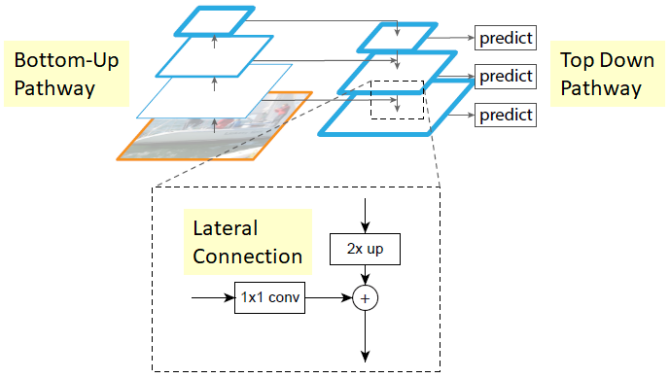


Fig. 3: The FPN architecture.

FPN comprises a backbone network, such as ResNet, which generates feature maps at various scales, and a pyramid network that consolidates these feature maps into a single feature map. The pyramid network includes lateral connections that link the feature maps at different scales and top-down connections that refine the feature maps at each scale. The final output of the network is a set of feature maps at multiple scales, which are utilized for object detection.

FPN has been demonstrated to outperform other state-of-the-art object detection models in terms of accuracy and speed. It has also been effectively utilized in other computer vision tasks, such as semantic segmentation and instance segmentation, leading to remarkable results.

4) PSPNet: PSPNet, or Pyramid Scene Parsing Network [22], is a deep learning architecture used for semantic segmentation of images and has achieved state-of-the-art results on various benchmark datasets. The main idea behind PSPNet is to utilize the global context information of the image to improve the segmentation accuracy. PSPNet does this by exploiting a pyramid pooling module that aggregates features

from different levels of the network, capturing global context information at multiple scales.

The PSPNet architecture consists of a backbone network, such as ResNet or VGG, followed by a pyramid pooling module and a decoder network. The pyramid pooling module is the key component of the PSPNet and is used to capture global context information from the image. It takes feature maps from different levels of the backbone network and performs pooling operations at multiple scales, producing feature maps with different resolutions. These feature maps are then concatenated and fed into the decoder network to produce the final segmentation map.

The pyramid pooling module in PSPNet has four pooling operations with different window sizes, capturing global context information at different scales. The feature maps generated by these pooling operations are then upsampled and concatenated to form a multi-scale feature map that is passed to the decoder network. The decoder network uses skip connections to fuse the high-resolution feature maps with the multi-scale feature map to produce the final segmentation map.

PSPNet has several advantages over other segmentation models. Firstly, it can capture global context information effectively, allowing it to handle images with large variations in scale and viewpoint. Secondly, it is computationally efficient since it does not require extra convolutional layers, making it suitable for real-time applications. Finally, it is a generic framework that can be applied to various tasks such as image classification and object detection.

PSPNet is a powerful deep learning architecture for semantic segmentation tasks, which utilizes a pyramid pooling module to capture global context information effectively. It has achieved state-of-the-art results on various benchmark datasets, and its generic framework allows it to be applied to various computer vision tasks.

5) MANet: MANet [23] consists of a backbone network that extracts feature maps at different scales, a multi-scale attention module, and a decoder network that produces the final segmentation mask. The multi-scale attention module has two components: a global attention module and a local attention module. The global attention module captures the global contextual information of the image, while the local attention module captures the local contextual information at different scales.

In the global attention module, a feature map is first fed into a global pooling layer to produce a global descriptor. This global descriptor is then passed through a fully connected layer and a sigmoid activation function to produce a global attention map. The global attention map is used to weigh the feature maps of all scales, emphasizing the most informative features for segmentation.

In the local attention module, feature maps at different scales are first fed into a pyramid pooling module to produce fixed-size feature maps. These fixed-size feature maps are then concatenated and passed through a convolutional layer and a sigmoid activation function to produce a local attention map. The local attention map is used to weigh the feature maps at each scale, allowing the network to focus on the most relevant features for segmentation.

Finally, the decoder network upsamples the feature maps to produce the final segmentation mask. The decoder network consists of several convolutional layers and skip connections that combine the feature maps of different scales to improve the accuracy of segmentation.

MANet has shown to achieve state-of-the-art results in various semantic segmentation benchmarks. It can handle complex scenes with objects of various sizes and shapes and is robust to occlusions and cluttered backgrounds. MANet's attention mechanisms allow it to focus on the most relevant features for segmentation, making it more efficient and accurate than other segmentation models.

6) LinkNet: LinkNet [24] is a deep learning architecture designed for semantic segmentation, aiming to tackle challenges such as high computational costs and the need for large amounts of training data in large-scale image segmentation tasks. Its main characteristic is a lightweight and computationally efficient architecture, composed of blocks containing a downsampling layer, an encoding layer with residual connections, and an upsampling layer that increases spatial resolution.

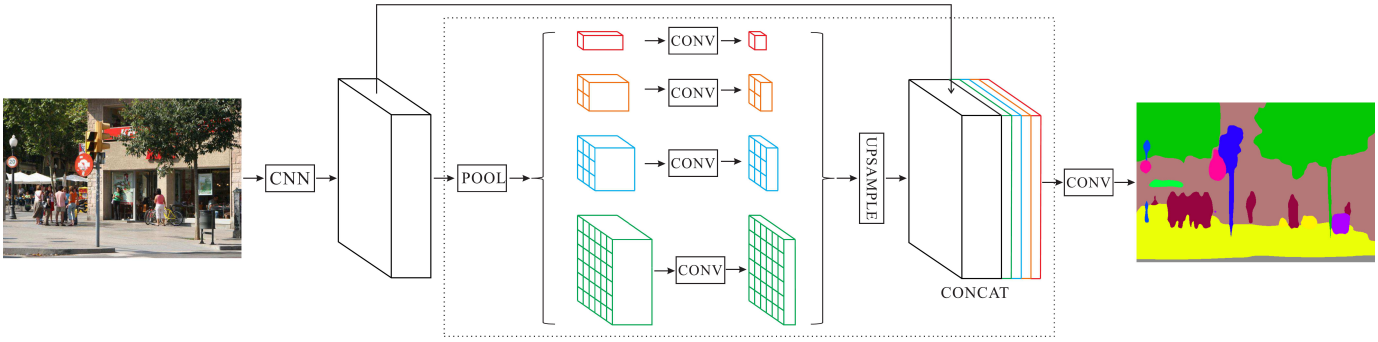
One of LinkNet's strengths is its use of skip connections, which enable the incorporation of features from different scales into the final segmentation map. These connections are placed between the encoding and upsampling layers, allowing for the integration of both low-level and high-level features in the final segmentation map.

LinkNet has demonstrated superior performance over other lightweight models across various benchmark datasets and has been applied to different image segmentation tasks, including medical imaging and aerial image segmentation. Its ability to handle large-scale segmentation tasks with a smaller computational footprint and less training data makes it a valuable tool in the field of computer vision.

B. Thresholding

1) Isodata: The Isodata thresholding technique [25] is a classic image segmentation method that separates the foreground and background regions of an image by thresholding its intensity values. The method aims to minimize the variance within each region by finding an optimal threshold value. Initially, the image is divided into two regions based on an initial threshold value. The mean intensity values of the two regions are then computed to calculate a new threshold value. This iterative process is continued until the threshold value converges or a maximum number of iterations is reached. The final threshold value is used to segment the image into two regions. The Isodata thresholding technique is simple and effective for images with non-uniform background intensity. However, it may be sensitive to the initial threshold value and affected by image noise. Moreover, it may not be suitable for images with multiple objects or complex intensity distributions.

2) Li: The Li thresholding technique, proposed by Li et al. [26], is an effective image segmentation method that leverages thresholding of the image intensity values to separate the foreground and background regions. The primary objective of the method is to determine a threshold value that can divide the



(a) Input Image (b) Feature Map (c) Pyramid Pooling Module (d) Final Prediction

Fig. 4: The PSPNet architecture.

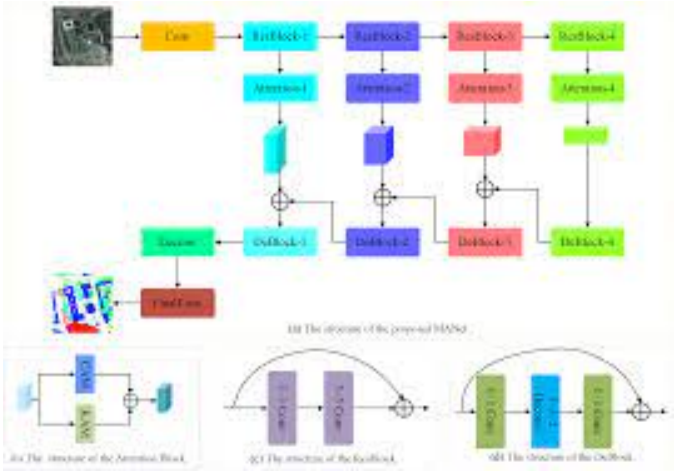


Fig. 5: The MANet architecture.

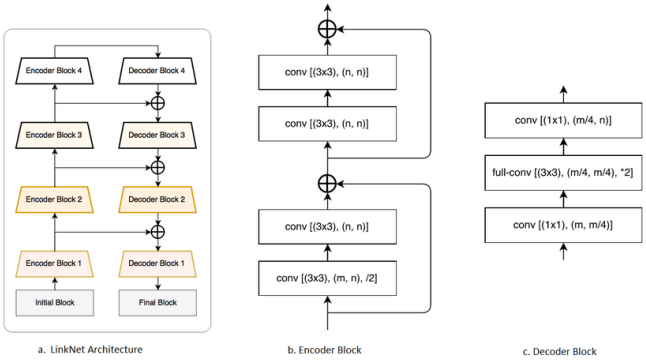


Fig. 6: The LinkNet architecture.

image into two regions in such a way that the total variances within each region are minimized.

To perform image segmentation using the Li thresholding technique, an initial threshold value is used to divide the image into two regions. Next, the variance of intensity values within each region is calculated, and the threshold value is optimized by minimizing the sum of variances between the two regions. This iterative process is continued until the threshold value converges or the maximum number of iterations is reached. Finally, the optimized threshold value is used to segment the image into the foreground and background regions.

The Li thresholding technique offers various advantages, such as its ability to handle images with non-uniform background intensity and robustness to noise. Additionally, the method is well-suited for images with multiple objects or complex intensity distributions since it takes into account both the mean and variance of intensity values within each region.

However, the Li thresholding technique also has some limitations. The method requires a large number of iterations to converge, which can be computationally expensive. Additionally, it is sensitive to the choice of initial threshold value and can be affected by the presence of outliers or extreme intensity values in the image.

3) Mean: The Mean thresholding technique [27] is a simple method for image segmentation that is based on the idea of thresholding the image intensity values to separate the foreground and background regions. The goal of the method is to find a threshold value that separates the image into two regions such that the mean intensity values within each region are distinct. The Mean thresholding technique starts by computing the mean intensity value of the entire image. This mean value is then used as the threshold value to segment the image into two regions: the foreground and background. The foreground region is defined as the region of the image with intensity values greater than the mean, while the background region is defined as the region with intensity values less than or equal to the mean.

The Mean thresholding technique is simple to implement and can be used for images with a clear distinction between the foreground and background regions. However, it is not well suited for images with multiple objects or complex intensity distributions, as it does not consider the variance of the intensity values within each region.

Additionally, the Mean thresholding technique can be affected by noise in the image and may not produce accurate results in such cases. In order to overcome this limitation, more sophisticated thresholding techniques, such as Otsu's method or the Li thresholding technique, can be used.

4) Minimum: The Minimum thresholding technique [27] is an image segmentation method that involves thresholding the image intensity values to distinguish the foreground and background regions. Its objective is to determine a threshold value that splits the image into two regions, each with distinctive intensity values.

The Minimum thresholding technique commences with identifying the minimum intensity value in the image. This value is then employed as the threshold value to separate the image into two regions: the foreground and the background. The foreground region is defined as the region of the image where intensity values are greater than the minimum, while the background region corresponds to the region where the intensity values are equal to the minimum.

The Minimum thresholding technique is a straightforward method for image segmentation, yet it has several shortcomings. The technique is unsuitable for images that contain multiple objects or have complex intensity distributions, as it does not factor in the mean or variance of the intensity values in each region. Moreover, the method can be influenced by noise in the image, which can lead to inaccurate outcomes.

The Minimum thresholding technique is restricted to binary image segmentation since it segments the image into only two regions - foreground and background - based on a single threshold value. To segment images into more than two regions, advanced thresholding methods like multi-level thresholding or region-based thresholding are necessary.

5) Otsu: The Otsu thresholding technique [28] is a popular image segmentation method that relies on thresholding the intensity values of an image to differentiate between the foreground and background regions. The objective is to find a threshold value that separates the image into two regions, such that the variance between the intensity values within each region is maximized.

The Otsu method commences by computing the histogram of the image intensity values, which is then utilized to estimate the probability density function (PDF) of the image. The threshold value is then determined by maximizing the variance between the two regions defined by the threshold. The variance between the two regions is calculated as the weighted sum of the variances of the foreground and background regions, where the weights are proportional to the proportion of pixels in each region. The algorithm has been described in Algorithm 1.

Algorithm 1 Otsu Algorithm

procedure OTSU(X)
    Compute histogram (H) and probabilities (P_I) of each intensity level (I) for input image (X).
    Initialize: $\omega_i(0), \mu_i(0)$
    for each threshold t in $[1, max(I)]$ **do**
        Update: ω_i, μ_i
        Compute: $\sigma_b^2(t)$
    end for
    Return $\underset{t}{\operatorname{argmax}}\{\sigma_b^2(t)\}$
end procedure

The Otsu thresholding technique is most effective for images that have bimodal intensity distributions, where the foreground and background regions have distinct intensity values. The method is robust to noise and can handle images with complex intensity distributions. Additionally, it is efficient and can be implemented using a single pass of the image histogram.

However, the Otsu thresholding technique has its limitations. The method assumes that the image intensity values are Gaussian-distributed, and may not provide accurate results for images with non-Gaussian intensity distributions. Furthermore, the method can only be used for binary image segmentation, as the foreground and background regions are defined by a single threshold value. To segment images into more than two regions, more sophisticated thresholding techniques, such as multi-level thresholding or region-based thresholding, must be employed.

6) Triangle: The Triangle Thresholding [29] technique is a widely-used and effective image thresholding method in computer vision and image processing. It relies on histogram analysis, which visually represents the distribution of pixel intensities in an image.

To perform image segmentation, the technique divides the histogram into two parts: one part represents the background, while the other part represents the object of interest. The threshold value is then calculated as the point where the maximum variance occurs between these two parts.

To find the threshold value, the algorithm calculates the histogram of the image and identifies the two peaks, one representing the background and the other representing the object of interest. The threshold value is then determined by minimizing the sum of variances of the two peaks, also known as the threshold value. After obtaining the threshold value, the image is converted into a binary image. Pixels with intensity values greater than the threshold value are set to 1 (foreground), while pixels with intensity values less than the threshold value are set to 0 (background).

The Triangle Thresholding technique is efficient and straightforward, making it ideal for real-time applications. It is also robust against noise and illumination changes, making it a popular choice for image thresholding. Furthermore, it does not require any prior knowledge about the image, making it a suitable technique for various thresholding applications.

7) Yen: Yen thresholding technique [30] is a popular image segmentation method used in digital image processing. It is based on the idea of finding the optimal threshold value that separates the image into meaningful regions while minimizing the information loss. The technique has been widely used in various applications such as medical imaging, document analysis, and computer vision.

The idea behind the Yen thresholding technique is to find a threshold value that maximizes the between-class variance while minimizing the within-class variance. In other words, it tries to find the threshold value that separates the image into two regions with maximum separability while keeping the homogeneity of each region intact. This is achieved by finding the threshold that minimizes the information entropy of the two resulting regions. The Yen thresholding technique starts by calculating the cumulative histogram of the image intensity values.

Next, it searches for the threshold value that minimizes the information entropy by considering all possible threshold values. The entropy of a region is calculated by summing up the probabilities of each intensity level in the region, and taking the negative logarithm of the sum. The threshold

value that minimizes the entropy is considered as the optimal threshold for the image.

One of the advantages of the Yen thresholding technique is its ability to handle images with non-uniform background intensity values and multiple peaks in the histogram. It can also handle images with noise and outliers by considering the overall distribution of intensity values in the image.

C. Other Methods

1) Region Filling: Image processing commonly uses region filling as a segmentation technique to isolate objects or areas of interest in an image. The process begins by selecting a seed point and gradually expanding the region by adding adjacent pixels that meet specific criteria. Depending on the image and its intended application, the criteria for adding neighboring pixels may vary. Simple intensity values may be sufficient in some cases, while more intricate criteria such as gradient direction or texture may be required in others.

Despite its effectiveness as a simple and computationally efficient method for image segmentation, region filling can be sensitive to the choice of seed point and may face challenges when dealing with complex images. Nonetheless, region filling remains a valuable tool in image processing and has been utilized in various fields such as medical imaging, remote sensing, and quality control. With the advancements in machine learning and computer vision, more sophisticated region filling algorithms have been developed to handle intricate images and address the limitations of traditional approaches

2) Chan-Vese Segmentation: The Chan-Vese segmentation [31] technique is a popular method in image processing for separating objects from the background in digital images. It uses the level set method to evolve a contour that separates the object of interest from the background.

The Chan-Vese functional minimizes a functional that measures the similarity between the object and the background while promoting smoothness of the contour. The functional has two terms, one measuring the difference in intensity between the object and the background, and the other measuring the smoothness of the contour. The Chan-Vese method does not require prior knowledge of the shape or size of the object to be segmented; it relies solely on the intensity differences between the object and the background to segment the image.

The Chan-Vese method has been successfully applied in various fields such as medical image analysis, object recognition, and image segmentation. Furthermore, it has been extended to handle more complex image structures, including multiple objects, and combined with other techniques such as active contours and level set methods to improve its performance.

IV. Experimental Results

In this section we analyse the performances of each method discussed quantitatively and exhaustively. Details of the dataset have been provided beforehand. We provide a detailed description of the evaluation metrics used as well.

A. Evaluation metrics

We use mean Intersection-over-Union score for evaluating segmentation performance with respect to ground-truths. For a given image,X, let the segmentation output is given as $\hat{Y}$ and ground truth as Y. Then, the IOU score is given as:

$$IOU = \frac{Y \cdot \hat{Y}}{Y + \hat{Y} + \epsilon} \tag{1}$$

ϵ is a very small value (typically 10^{-5}) used to avoid floating point errors due to division by zero when the denominator of the fraction becomes zero. If IOU score for the i^{th} sample in a dataset of N images is IOU_i, then mean IOU score ($mIOU$) is given as:

$$mIOU = \frac{1}{N} \sum_{i=1}^{N} IOU_i \tag{2}$$

All results reported are $mIOU$ scores computed on the test set.

B. Various segmentation methods

We have tested several deep learning based methods in Table I and non-deep learning based methods in Table II. We have also compared these methods Visually in Figure 7 and Figure 8. There are mainly three distinctions:

1) Deep Learning Based Methods
2) Thresholding Methods
3) Other Methods

1) Deep Learning Based Methods: The deeplearning methods namely tested were, UNet, UNet++, FPN, PSPNet, MANet, LinkNet. There trainings have been performed for with and without noise corruption as tabulated in Table I. The severity of noise corruption has been done on 3 levels. The choices for noise corruptions are Gaussian, Salt and Pepper, Speckle and Poisson where normal means without any corruption.

2) Thresholding Methods: The study experimented with several thresholding-based image segmentation techniques, as shown in Table II. Noise corruptions were not used because they were deemed to be counterproductive to the objective of these methods. Figure 8 provides a visualization of the segmentation masks produced by the thresholding-based techniques. Based on the results, the Otsu thresholding technique was found to be the most effective among the thresholding techniques used.

3) Other Methods: In this section we get the experimental results on two methods, region filling, Chan Vese as shown in Table II. The visualisations are provided in Figure 8.

V. Conclusion

Data plays a key role in the performance of these models even though immense research efforts are being employed. Uncertainty of datasets will inevitably lead to models performing poorly if it is too prevalent in the dataset. Our experimental results show that the results obtained from the thresholding

TABLE I: $mIOU$ scores on Test Set for Deep Learning based methods (for each model, scores on the left pertains to the model trained without noisy data, and the score on the right pertains to the model trained with noisy data).

Data (Severity)		***UNet***		***Unet++***		***FPN***		***PSPNet***		***LinkNet***		***MANet***	
Normal		0.8932	0.9054	0.9042	0.9046	0.8937	0.8968	0.8931	0.8994	0.8979	0.9050	0.8972	0.8894
Gaussian Noise	**1**	0.8934	0.9051	0.9034	0.9044	0.8934	0.8962	0.8940	0.8992	0.8979	0.9049	0.8970	0.8887
	2	0.8936	0.9058	0.9043	0.9048	0.8935	0.8949	0.8935	0.8990	0.8973	0.9047	0.8965	0.8888
	3	0.8921	0.9049	0.9033	0.9039	0.8934	0.8975	0.8938	0.8992	0.8981	0.9048	0.8960	0.8879
Salt and Pepper Noise	**1**	0.6194	0.9010	0.8330	0.8920	0.8417	0.8894	0.8402	0.8966	0.8083	0.9013	0.7974	0.8740
	2	0.6824	0.9009	0.8257	0.8936	0.8578	0.8821	0.8408	0.8946	0.7985	0.9004	0.8130	0.8685
	3	0.6565	0.9015	0.8311	0.8917	0.8575	0.8870	0.8225	0.8963	0.8086	0.9003	0.7870	0.8824
Speckle Noise		0.0004	0.0014	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0011	0.0002	0.0005	0.0001
Poisson Noise		0.8940	0.9051	0.9042	0.9042	0.8934	0.8974	0.8935	0.8994	0.8979	0.9051	0.8971	0.8900

TABLE II: $mIOU$ scores on Test Set for Non-Deep Learning based methods

Thresholding Based Techniques							***Miscellaneous Techniques***	
Isodata	*Li*	*Mean*	*Min*	*Otsu*	*Triangle*	*Yen*	*Region Filling*	*Chan-Vese*
0.3635	0.3104	0.3046	0.2167	0.3646	0.1658	0.3548	0.1394	0.3058

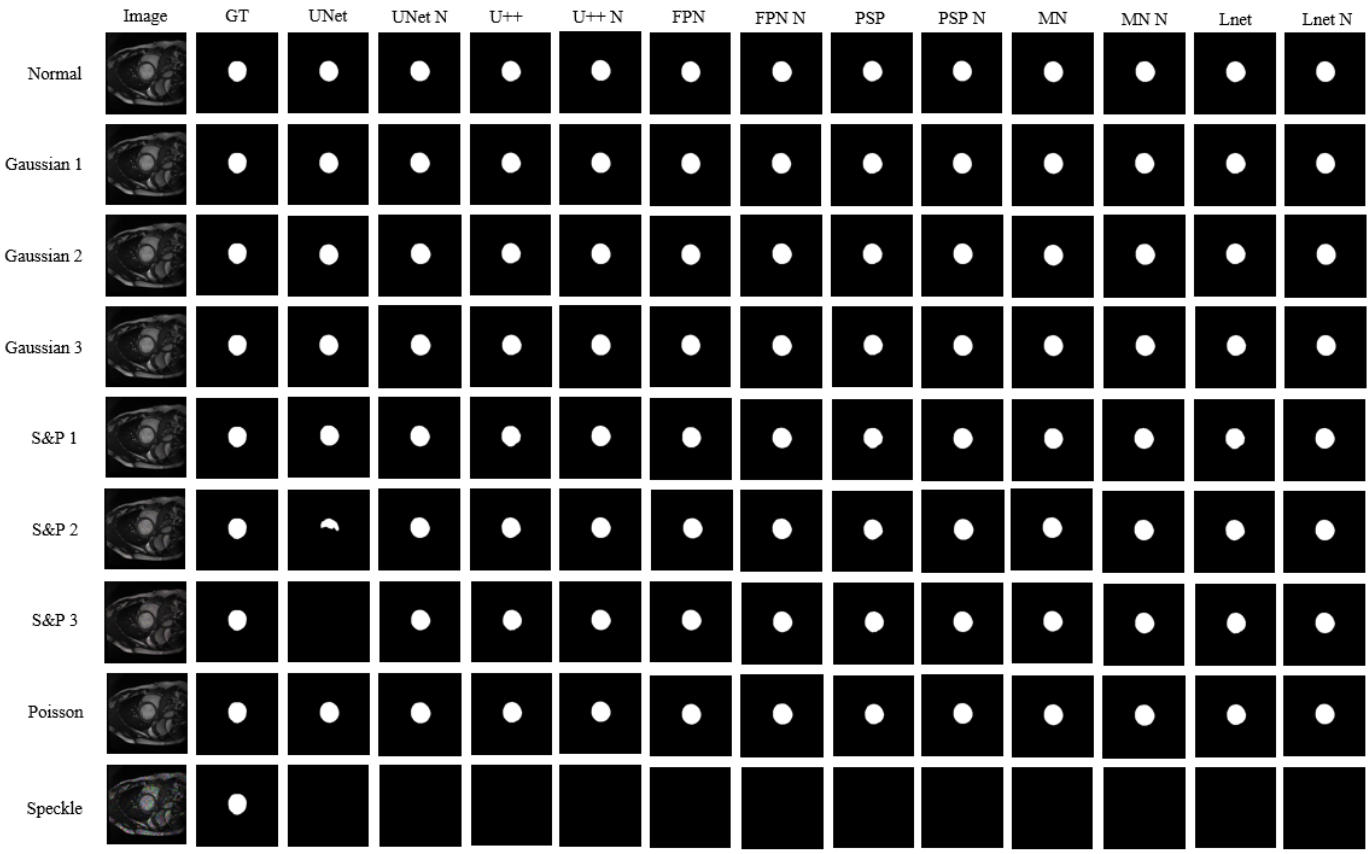


Fig. 7: A visual inspection of the segmentation masks generated by the various standard deep learning methods. **Note:** 'N' indicates model has been trained on noisy data. Absence of the character means the trianing was done on normal data. *Abbreviations*: U++: UNet++, FPN: Feature Pyramidal Networks, PSP: PSPNet, MN: MANet, Lnet: LinkNet

based schemes are generally comparable. However, segmentation performance improves significantly with the use of the deep learning based segmentation algorithms.

We also observe that the segmentation results are sufficiently good enough for deployment in commercial purposes or as a diagnosis tool in a medical environment. However, this is still not practical due to the computational complexity of deep learning and other algorithms and therefore, requires further work on making the pipeline more efficient without compromising on results.

REFERENCES

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[2] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[3] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[4] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[5] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[6] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*. PMLR, 2019, pp. 6105–6114.

[7] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer, 2015, pp. 234–241.

[8] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4*. Springer, 2018, pp. 3–11.

[9] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[10] J. Causey, J. Stubblefield, J. Qualls, J. Fowler, L. Cai, K. Walker, Y. Guan, and X. Huang, "An ensemble of u-net models for kidney tumor segmentation with ct images," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 19, no. 3, pp. 1387–1392, 2021.

[11] J. Li, Z. L. Yu, Z. Gu, H. Liu, and Y. Li, "Dilated-inception net: multi-scale feature aggregation for cardiac right ventricle segmentation," *IEEE Transactions on Biomedical Engineering*, vol. 66, no. 12, pp. 3499–3508, 2019.

[12] M. Tabassum, T. M. Khan, M. Arsalan, S. S. Naqvi, M. Ahmed, H. A. Madni, and J. Mirza, "Cded-net: Joint segmentation of optic disc and optic cup for glaucoma screening," *IEEE Access*, vol. 8, pp. 102 733–102 747, 2020.

[13] Q. Liu, X. Hong, S. Li, Z. Chen, G. Zhao, and B. Zou, "A spatial-aware joint optic disc and cup segmentation method," *Neurocomputing*, vol. 359, pp. 285–297, 2019.

[14] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang, "Swin-unet: Unet-like pure transformer for medical image segmentation," in *Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part III*. Springer, 2023, pp. 205–218.

[15] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "Transunet: Transformers make strong encoders for medical image segmentation," *arXiv preprint arXiv:2102.04306*, 2021.

[16] H.-Y. Zhou, J. Guo, Y. Zhang, L. Yu, L. Wang, and Y. Yu, "nnformer: Interleaved transformer for volumetric segmentation," *arXiv preprint arXiv:2109.03201*, 2021.

[17] N. Ibtehaz and M. S. Rahman, "Multiresunet: Rethinking the u-net architecture for multimodal biomedical image segmentation," *Neural networks*, vol. 121, pp. 74–87, 2020.

[18] R. Azad, A. Bozorgpour, M. Asadi-Aghbolaghi, D. Merhof, and S. Escalera, "Deep frequency re-calibration u-net for medical image segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 3274–3283.

[19] P. Tang, X. Yan, Q. Liang, and D. Zhang, "Afln-dgcl: Adaptive feature learning network with difficulty-guided curriculum learning for skin lesion segmentation," *Applied Soft Computing*, vol. 110, p. 107656, 2021.

[20] O. Bernard, A. Lalande, C. Zotti, F. Cervenansky, X. Yang, P.-A. Heng, I. Cetin, K. Lekadir, O. Camara, M. A. G. Ballester, *et al.*, "Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: is the problem solved?" *IEEE transactions on medical imaging*, vol. 37, no. 11, pp. 2514–2525, 2018.

[21] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125.

[22] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2881–2890.

[23] T. Fan, G. Wang, Y. Li, and H. Wang, "Ma-net: A multi-scale attention network for liver and tumor segmentation," *IEEE Access*, vol. 8, pp. 179 656–179 665, 2020.

[24] A. Chaurasia and E. Culurciello, "Linknet: Exploiting encoder representations for efficient semantic segmentation," in *2017 IEEE visual communications and image processing (VCIP)*. IEEE, 2017, pp. 1–4.

[25] F. R. Velasco, "Thresholding using the isodata clustering algorithm," MARYLAND UNIV COLLEGE PARK COMPUTER SCIENCE CENTER, Tech. Rep., 1979.

[26] C. H. Li and C. Lee, "Minimum cross entropy thresholding," *Pattern recognition*, vol. 26, no. 4, pp. 617–625, 1993.

[27] C. A. Glasbey, "An analysis of histogram-based thresholding algorithms," *CVGIP: Graphical models and image processing*, vol. 55, no. 6, pp. 532–537, 1993.

[28] D. Liu and J. Yu, "Otsu method and k-means," in *2009 Ninth International conference on hybrid intelligent systems*, vol. 1. IEEE, 2009, pp. 344–349.

[29] G. W. Zack, W. E. Rogers, and S. A. Latt, "Automatic measurement of sister chromatid exchange frequency." *Journal of Histochemistry & Cytochemistry*, vol. 25, no. 7, pp. 741–753, 1977.

[30] J.-C. Yen, F.-J. Chang, and S. Chang, "A new criterion for automatic multilevel thresholding," *IEEE Transactions on Image Processing*, vol. 4, no. 3, pp. 370–378, 1995.

[31] T. Chan and L. Vese, "An active contour model without edges," in *Scale-Space Theories in Computer Vision: Second International Conference, Scale-Space'99 Corfu, Greece, September 26–27, 1999 Proceedings 2*. Springer, 1999, pp. 141–151.

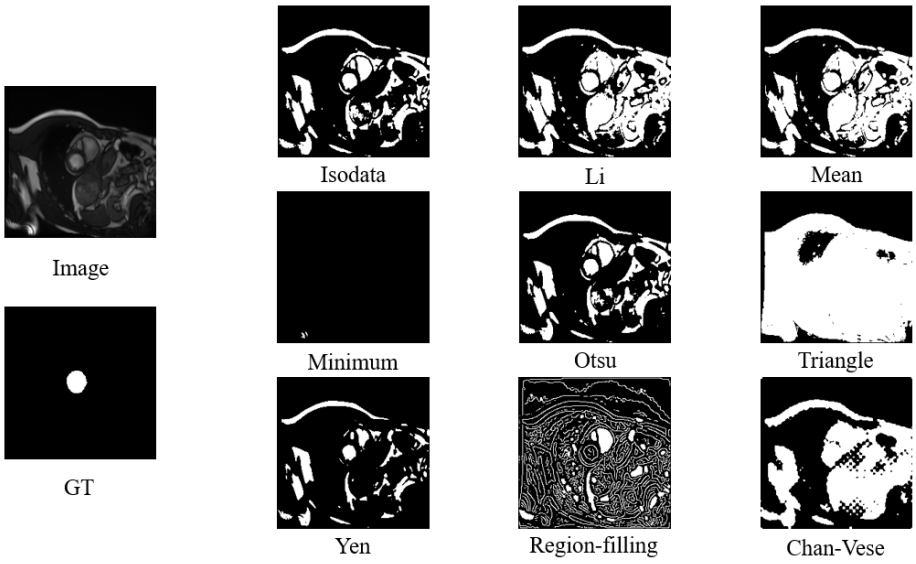


Fig. 8: A visual inspection of the segmentation masks generated by the various standard non-deep learning methods.