EM-Net: An Efficient M-Net for segmentation of surgical instruments in colonoscopy frames

Debapriya Banik¹ ², Kaushiki Roy¹ ², and Debotosh Bhattacharjee¹

1. Department of Computer Science and Engineering, Jadavpur University, Kolkata, India
2. Both shares Co-first Authorship

Abstract
This paper addresses the Instrument Segmentation Task, a subtask for the “MedAI: Transparency in Medical Image Segmentation” challenge. To accomplish the subtask, our team “Med_Seg_JU” has proposed a deep learning-based framework, namely “EM-Net: An Efficient M-Net for segmentation of surgical instruments in colonoscopy frames”. The proposed framework is inspired by the M-Net architecture. In this architecture, we have incorporated the EfficientNet B3 module with U-Net as the backbone. Our proposed method obtained a Jaccard Coefficient of 0.8205, Dice Similarity Coefficient of 0.8632, Precision of 0.9005, F1-score of 0.8632, and Pixel Accuracy of 0.9799 as evaluated by the challenge organizers on a separate test dataset. These results justify the efficacy of our proposed method in the segmentation of the surgical instruments.

Keywords: Deep learning; Segmentation; Colonoscopy; Surgical instruments

Introduction
The automated and robust segmentation of gastrointestinal (GI) instruments plays a vital role in the success of minimally invasive surgeries [1]. The instrument segmentation aids the doctors or surgical robots to overcome the complexities during performing lower GI endoscopy (colonoscopy) or upper GI endoscopy (gastroscopy) surgeries specifying the precise location, orientation, and current status of the instrument [2]. However, the automated segmentation of GI instruments from the colonoscopy or gastroscopy frames is challenging since the instruments operate in an uncontrolled environment and suffer from illumination changes, mirror reflections. Moreover, the contour separating the instrument and the organ in contact is weak [3]. The subtask provided for the challenge “MedAI: Transparency in Medical Image Segmentation” is to develop computer-aided techniques for automatic segmentation of surgical tools with a high degree of segmentation metrics [4]. So to accomplish this task, we have proposed a deep learning framework, namely “EM-Net: An Efficient M-Net for segmentation of surgical instruments in colonoscopy frames”.

Materials and methods
Preprocessing
The training challenge dataset [5] contains a limited number of frames with surgical instruments. So to enhance the size, different augmentation techniques such as random rotation (30° to 330°), flipping (vertical and horizontal) [6] are employed before data is fed into the network.

Proposed EM-Net Architecture
Our proposed EM-Net is inspired by the M-Net architecture [7], an encoder-decoder-based architecture as shown in Figure 1. The encoder branch uses EfficientNet B3 [8] as the backbone. At each level of the encoder branch, a 2D convolution layer (same padding) followed by batch normalization and ReLU activation function is employed. A multiscale input is provided in the network where the input image is downsampled at a rate (2, 4, and 8) at each level of the encoder branch and fed into the network to provide a multi-level receptive field. An average pooling operation is used to downsample the feature maps in the encoder branch. The decoder branch is a mirror structure to the encoder branch, where an Upsampling is used to downsample the feature maps in each level of the decoder. A skip connection is used from each level of the encoder to the corresponding level of the decoder to enhance the flow of spatial information lost during downsampling. The final feature maps from each level of the

© 2021 Author(s). This is an open access article licensed under the Creative Commons Attribution License 4.0. (http://creativecommons.org/licenses/by/4.0/).
decoder branch undergo point-wise convolution followed by sigmoid activation, and the output is upsampled as the size of the final output image to provide deep supervision and provide a local pixel-level prediction map for each scale of the input image. Finally, all of them are fused to generate the segmentation mask.

Post-Processing
The output predicted mask from the EM-Net contains some minor false-positive regions, as shown in Figure 2b. So we ignored such regions by discarding the connected pixel region with pixels less than 300, followed by morphological hole filling operations. Finally, the convex hull generates the final segmentation mask, as shown in Figure 2c.

Implementation Details
The Dice loss function is used to optimize and learn the objective of the proposed network. The Adam optimizer with an initial learning rate of 0.0001 is used, and the rate is updated every 10 epochs. The model is trained for 300 epochs with a batch size of 12.

Results and Discussion
Database Description
We were provided with a dataset, "Kvasir-Instrument," having 590 frames with corresponding ground-truth [5]. Before the training process, the images and the respective ground-truth masks are resized to $256 \times 256$. The dataset is randomly split into 70:30 for training and validation. For testing, the challenge organizers provided a separate test set containing 300 images without ground truth.

Table 1: Official Results of our approach on test dataset as evaluated by the challenge organizers.

<table>
<thead>
<tr>
<th>Model</th>
<th>JC</th>
<th>DSC</th>
<th>PRE</th>
<th>REC</th>
<th>F1</th>
<th>ACC</th>
</tr>
</thead>
<tbody>
<tr>
<td>EM-Net</td>
<td>0.8205</td>
<td>0.8632</td>
<td>0.8632</td>
<td>0.9005</td>
<td>0.8632</td>
<td>0.9799</td>
</tr>
</tbody>
</table>

Qualitative and Quantitative evaluation
In this subsection, we present the results obtained on the test dataset provided by the challenge organizers. In the 1\textsuperscript{st} row of Figure 3, we have considered some frames from the test dataset showing different complex variations of surgical instruments in terms of shapes and sizes. The corresponding segmentation mask obtained by our proposed method is shown in the 2\textsuperscript{nd} row. It can be well perceived that our proposed method can efficiently segment the surgical instruments from the frames. In Table 1, we have tabulated the official results as evaluated by the challenge organizers on the test dataset. To evaluate the performance of the proposed method, different standard performance metrics are considered, such as Jaccard Coefficient (JC), Dice Similarity Coefficient (DSC), Precision (PRE), Recall (REC), F1-Score (F1), and Pixel Accuracy (ACC). From Table 1, it can be observed that the results are robust and consistent.

Conclusion and future work
This study addresses the surgical instrument segmentation task, a subtask in the "MedAI: Transparency in Medical Image Segmentation" challenge. To accomplish the task, we have proposed a deep learning-based framework, namely "EM-Net", to effectively segment the surgical instruments from the colonoscopy frames. Although our proposed method shows consistency in the segmentation of the surgical instruments as evaluated on a separate hidden test set. However, there were some frames where the poor lighting conditions and artifacts affected the surgical instruments, and in such situations, our proposed method failed to segment the instruments. As a part of future work, we will explore some higher versions of the efficient nets and study the generalization proficiency of the proposed network by exploring other medical datasets to justify its robustness.
References