Improving Loss Function for a Deep Neural Network for Lesion Segmentation

Bao Anh Trinh
baoanhmta@gmail.com
Le Quy Don Technical University
Hanoi, Vietnam

Trinh Thi Thuy An
anttt@mta.edu.vn
Le Quy Don Technical University
Hanoi, Vietnam

Ly Vu
vu.ly@lqdtu.edu.vn
Le Quy Don Technical University
Hanoi, Vietnam

Hang Dao
daoviethang@hmu.edu.vn
Hanoi Medical University
Hanoi, Vietnam

Thuy Nguyen
thuy.nguyen43@rmit.edu.vn
RMIT University
Ho Chi Minh city, Vietnam

ABSTRACT

Identifying and segmenting lesions are challenging tasks in automatic analysis of endoscopic images in computer aided diagnosis systems. Models based on encoder-decoder architectures have been proposed to segment lesions with promising results. However, those approaches have limitations in modeling the local appearance, dealing with imbalanced data, and over-fitting. This paper proposes a novel method to address these limitations. We propose to improve a state-of-the-art encoder-decoder based model for image segmentation by introducing a new loss function for training. The novel loss function is called Focal - Binary Cross Entropy - Intersection over Union (FBI), consisting three terms, a Focal term, a Binary Cross Entropy (BCE) term, and an Intersection over Union (IoU) term. In addition, we employ the Lasso regularization sparsity technique in learning to reduce over-fitting. As a result, the proposed model can effectively segment lesions of various sizes and shapes, leading to improved accuracy of the lesion segmentation task. Our proposed model outperforms existing deep learning models on two challenging data sets of gastrointestinal endoscopy for cancerous lesion segmentation.

KEYWORDS

Medical image segmentation, deep learning, encoder-decoder network, Loss function, endoscopy

ACM Reference Format:
Bao Anh Trinh, Trinh Thi Thuy An, Ly Vu, Hang Dao, and Thuy Nguyen. 2023. Improving Loss Function for a Deep Neural Network for Lesion Segmentation. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (SoICT '23).* ACM, New York, NY, USA, 8 pages. https://doi.org/XXXXXXX.XXXXXXX

SoICT '23, June 03–05, 2018, Ho Chi Minh city, Vietnam

ACM ISBN 978-1-4503-XXXX-X/12/09…$15.00
https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Gastrointesinal (GI) lesions can occur in various parts of the GI tract. Some types of lesions can be precursors to dangerous cancers such as gastric cancer and colorectal cancer that result in over 640,000 deaths annually [3]. Lesion detection is often carried out through endoscopy, and diagnosis is carried out based on the examination of these images. However, endoscopic procedures may miss 20% to 47% of lesions [19] due to subjective and objective factors, such as equipment limitations, endoscopic techniques, and physician interpretation [18]. This underscores the importance of developing computer-assisted lesion segmentation systems that can aid physicians to reduce the risk of missing lesions during procedures.

Using deep learning for lesion segmentation has achieved exceptional accuracy on benchmark datasets [6, 8, 24, 36]. Among various deep learning based methods, the encoder-decoder architecture of Unet [36], which combines low-level concrete features and high-level abstract features of an input image, has proven to be one of the most effective architectures for the segmentation task. Variants of Unet such as UNet++ [51], ResUNet++ [13], and DoubleUNet [12] have improved upon the original Unet by adopting nested or stacking approaches. Recently, the ESFPNet architecture [6], which uses the Mix Transformer (MiT) encoder as the backbone and an efficient stage-wise feature pyramid (ESFP) as the decoder, is proved to be one of the most effective models for the segmentation task [21, 48, 50].

However, deep learning models in general and ESFPNet in particular for lesion segmentation still pose several limitations. First, the lesion areas usually have various textures, shapes, and sizes, which makes them difficult to recognize. Second, the lesion segmentation is formulated as a classification problem at the pixel level where every image pixel needs to be classified into a normal class or abnormal class. In which, the imbalance between number of lesion pixels and normal pixels of small-sized lesion area reduces the effectiveness of deep learning models. A number of loss functions have been proposed to deal with these problems when training deep learning models. However, some limitations remain.

To handle above challenging issues, this paper proposes an approach for improving training an encoder-decoder model for segmentation. In particular, we employ the ESFPNet model [6] and propose a new hybrid loss function for training it. The new loss function is a combination of a Focal loss, a Binary Cross Entropy

(BCE) loss, and a Intersection over Union (IoU) loss namely Focal-BCE-IoU (FBI) loss. The BCE loss is a popular loss function for image segmentation [1] that measures the difference between the predicted and ground truth masks by computing the cross-entropy. Besides, the IoU loss is suitable for evaluating the overlap or similarity between predicted segmentation masks and ground truth masks. However, the BCE and IoU loss may result in sub-optimal results when models trained with imbalance data [20]. Therefore, we combine Focal loss [22] with these loss functions to address the class or region imbalance and improve the overall performance of the segmentation model.

By using the proposed FBI loss function the ESFPNet model is able to learn features from multi-sized lesions while dealing with imbalanced data more effectively. Moreover, we propose using the Lasso regularization [44] as a regularization constraint to perform sparse in training and then prune the network. The experimental results show that our proposed method can enhance the overall accuracy for the lesion segmentation.

The main contributions of this paper are:

- To propose a new loss function called FBI (Focal-BCE-IoU) and a learning strategy based on Lasso regularization for training deep learning models for medical image segmentation.
- To propose using ESPFNet trained with FBI for lesions segmentation in gastrointestinal endoscopic images
- To describe two new data sets for lesion segmentation segmentation and cancer detection and a comprehensive set of experiments and evaluations of the proposed method on the two data sets.

The rest of this paper is structured as follows. Section 2 presents related work on lesion segmentation. Section 3 explains the backbone of the encoder-decoder based model. In Section 4, our proposed architecture, i.e., ESFPNet along with our proposed FBI loss and sparse technique, is described. The datasets, evaluation metrics, and experiment settings are described in Section 5, while Section 6 presents the evaluation results. The conclusion is presented in Section 7.

2 RELATED WORK

Recently, the deep learning based approaches are most widely used for medical image segmentation [8, 24, 36]. In particular, the U-Net architecture [36] is one of the most widely used for this task.

Many variants of U-Net [36] have been proposed to enhance its performance. For example, He et al. suggested Nested U-Net [11] which extends Unet by adding multiple nested pathways to capture features at different scales. Besides that, Schlemper et al. proposed Attention U-Net [40] which includes a self-attention mechanism to concentrate on relevant features while disregarding irrelevant ones. Alom et al. proposed R2U-Net [2], which is a recurrent residual convolutional neural network based on U-Net, to process with sequential data (e.g., videos). ResUnet++ [13] is a more advanced architecture for medical image segmentation that integrates residual connections and squeeze-and-excitation blocks into U-Net to improve gradient flow and optimize neural network models. Additionally, Multi-scale Spatial Feature Fusion Network (MSRF-Net) [41] was proposed as a deep learning-based model that employs an innovative Dual-Scale Dense Fusion (DSDF) block for both dual-scale feature exchange and a sub-network designed to exchange multi-scale features through the DSDF block to enhance the accuracy of the lesion segmentation problem.

With the robust development of transformer-based techniques in computer vision, there has been a growing number of models leveraging them to tackle segmentation tasks [6, 9, 38]. Some notable examples such as the FCN-Transformer model [38], i.e., a new deep learning architecture for lesion segmentation in colonoscopy images. This combines the advantages of Fully Convolutional Networks (FCN) [26] and Transformers to improve the segmentation accuracy; the FCB-SwinV2 Transformer [9], which are the combination of the FCB module and the Transformer module, results in the promising accuracy for the lesion segmentation problem.

Among them, ESFPNet [6] demonstrates superior segmentation performance compared to competing deep learning architectures. Furthermore, ESPFNet has the ability to produce near real-time results, leading to potential use in endoscopy examinations. Specifically, the study by Qi Chang et al. [6] has shown that ESFPNet-S has proven to be a viable method for real-time segmentation and detection of cervical lesions in Autofluorescence bronchoscopy (AFB) video segments. Therefore, we utilize this model as the backbone for our proposed lesion segmentation system.

Another important aspect of building a lesion segmentation model based on deep learning approach is designing of appropriate loss functions. The loss function aims to guide the training model to predict the mask that correctly distinguishes the pixels in and out of the lesion areas of the input image. The commonly loss functions are the BCE loss [43], the Dice loss [33], and the IoU loss [35]. However, these losses may result in sub-optimal results when dealing with class imbalance issues [20]. Thus, we propose the hyprid loss function of the Focal loss, the BCE loss and the IoU loss to enhance the accuracy of the lesion segmentation problem.

Weight decay is one of the most widely used regularization techniques in deep learning for improving generalization and robustness of neural network models [10]. Least Absolute Shrinkage and Selector Operator (Lasso) is a regularization technique used to minimize the regression coefficients of to regularize the model parameters [45]. Besides, it can be also used to reduce the model complexity of neural network models by minimize the weights of the neural network models. The model's weights are sparsed, leading to reducing the overfitting problem [4]. This technique is also used in several neural network models for handling overfitting problem [10, 23].

3 BACKGROUND

3.1 ESFPNet Architecture

ESFPNet uses the pretrained Mix Transformer (MiT) as an encoder as the backbone and Efficient Stage-wise Feature Pyramid (ESFP) structure as a decoder [6]. The MiT encoder, which is built based on the Vision Transformer (ViT) [49], combines diagonal pathways and self-attention across four stages. This approach provides both high-resolution coarse features and fine-grained features with lower resolution, and thus, enhancing semantic segmentation performance.

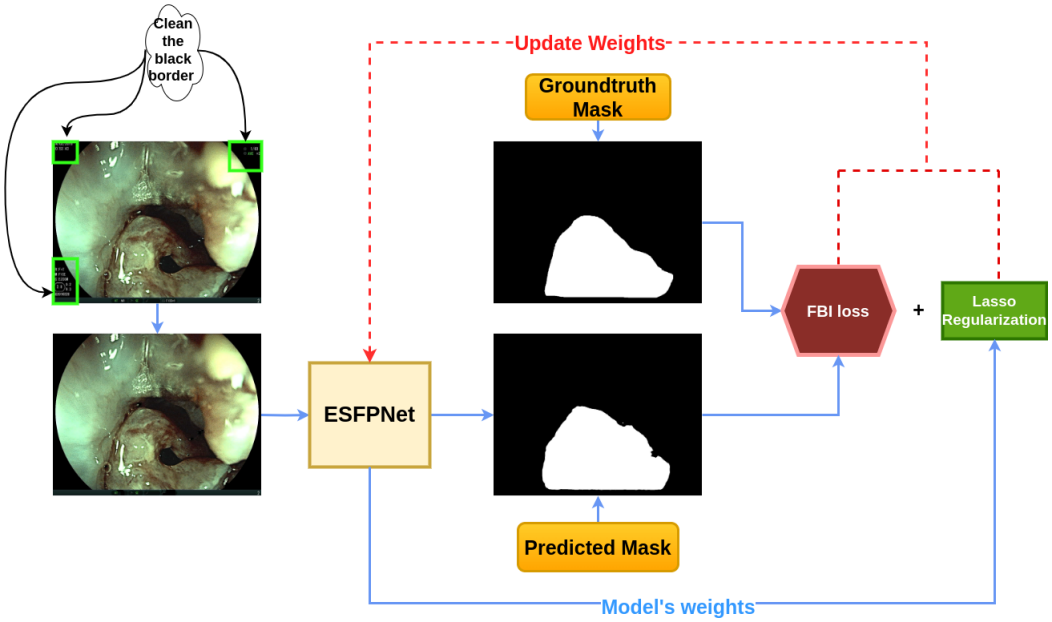

Figure 1: Architecture of Proposed Model for Lesion Segmentation.

To overcome the limitation of a small dataset, the pretrained MiT encoder is utilized and fine-tuned for specific tasks, outperforming Convolutional Neural Network models [48]. The ESFP decoder leverages multi-level features from the encoder, focusing on both local (shallow) and global (deep) features. Thus, it enhances the utilization of local features efficiently. Inspired by CfpNet [30], ESFP employs linear predictions for each stage's output to fuse them from global to local. These intermediate features of the decoder are then concatenated to generate the segmentation result.

3.2 Loss Function

Loss functions are vital component of deep neural networks when applied to image segmentation. They are used for quantifying the disparity between predicted and actual segmentation, guiding model optimization. Four common types of loss functions in image segmentation are the Distribution-Based loss function, the Region-Based loss function, the Boundary-Based loss function, and the Compound loss function [32].

The Distribution-Based loss focuses on matching pixels by considering the statistical distribution of pixel values in predicted and ground truth segmentation maps. This uses the cross-entropy to measure the difference of the distribution of pixel values. A number of loss function is calculated based on the cross-entropy, such as the TopK loss [47], the Distance map Penalized Cross Entropy (DPCE) loss [5].

The Region-based loss functions aim to minimize discrepancies or maximize the overlap of regions between the ground truth mask and the predicted segmentation mask. There are a number of these loss functions, i.e., the Dice loss [33], the IoU loss [35], the Tversky loss [37], and Penalty loss [42].

The Boundary-based loss leverages the difference of gradient or distance transforms to minimize the disparities between predicted and true boundaries. The Boundary loss [15] and the Hausdorff Distance loss [14] are examples of this type of loss function.

The Compound loss combines multiple components, addressing diverse segmentation quality aspects, yielding a comprehensive loss. It is useful for the image segmentation tasks requiring simultaneous optimization of accuracy, smoothness, and boundary preservation.

Selection of the appropriate loss assists the deep neural networks in learning the objective of the image segmentation problem effectively, leading to enhancing the performance and robustness of the image segmentation problem. In general, using a combined loss function yields better overall effectiveness compared to a single loss function [32].

4 PROPOSED METHOD

In this section, we introduce the proposed loss function, i.e., the FBI loss, and sparsity techniques, called as Lasso regularization in our training strategy to improve the performance of lesion segmentation. After that, we present the lesion segmentation model with the ESFPNet backbone trained by the proposed loss.

4.1 Focal-BCE-IoU Loss Function

The ESFPNet model [7] is trained by the combined loss function with two terms, i.e., BCE and IoU. However, the drawback of this loss function it is less effective with imbalanced datasets which include both large-sized and small-sized lesion areas. To handle the above issue of ESFPNet, we introduce the Focal-BCE-IoU (FBI) loss function which has three terms: BCE, IoU, and Focal loss term.

The IoU loss is a commonly used loss function in image segmentation tasks. It quantifies the dissimilarity between predicted and ground truth masks by measuring the spatial overlap between them and aims to maximize the intersection over union ratio, which is

an important metric for evaluating segmentation models. Here, we apply the weighted IoU loss is defined as following equation.

$$\mathcal{L}^{w}_{IoU} = 1 - \frac{\sum_{i=1}^{N}(we_i \times y_i \times p_i)}{\sum_{i=1}^{N} we_i \times (y_i + p_i - y_i \times p_i)}, \tag{1}$$

where N is the number of pixels in the image, y_i is value of pixel i in the ground truth mask [1], p_i is the predicted value of pixel i belonging to the lesion area. The weight matrix we is calculated based on the ground truth mask. It is the difference between the mask before and after smoothed by applying average pooling operation [6]. This aims to guide the loss focusing on more important regions during the training process [28, 29, 31].

The BCE loss is a common loss function used in image segmentation tasks to segment two class. The computation of this loss is based on cross-entropy which is used to measure the difference between two probability distributions. Thus, the BCE loss measures the differences in information content between two classes, e.g., the ground truth masks and predicted masks. It is effective with equal data-distribution between classes. We also apply the weighted BCE loss that is calculated as follows:

$$\mathcal{L}^{w}_{BCE} = -\frac{1}{\sum_{i=1}^{N} we_i} \sum_{i=1}^{N} we_i \times [y_i \log(p_i) + (1 - y_i) \log(1 - p_i)], \tag{2}$$

where the definitions of N, we, y_i, and p_i are similar to those in the IoU loss defined in Eq. 1.

The Focal loss is a specialized loss function that is introduced to address the issue of class imbalance due to the various sizes of lesion areas. This loss emphasizes the hard examples, e.g., the image with small sized lesion area, by assigning different weights to easy and hard examples. Specifically, in training a classifier, hard or misclassified examples are assigned higher weights while down-weighting easy examples. To focus on the important regions, we unitize the weighted Focal loss is defined as follows:

$$\begin{aligned}\mathcal{L}^{w}_{\text{Focal}} = -\frac{1}{\sum_{i=1}^{N} we_i} \sum_{i=1}^{N} we_i \times \big[& y_i (1 - p_i)^{\gamma} \log(p_i) \\ & + (1 - y_i) p_i^{\gamma} \log(1 - p_i),\end{aligned} \tag{3}$$

where N, we, y_i, and p_i are defined as in Eq. 1 , γ is a hyperparameter [2] controls the amount of reducing the loss for well-classified examples. A higher γ values put more emphasis on hard examples.

Our proposed loss function is the combination of the BCE, IoU, and Focal losses to enhance the lesion segmentation problem with both large and small-sized lesion areas. This computation is presented in Eq. 4.

$$\mathcal{L}_{FBI} = \frac{1}{N} \times (\mathcal{L}^{w}_{Focal} + \mathcal{L}^{w}_{BCE} + \mathcal{L}^{w}_{IoU}), \tag{4}$$

where $\mathcal{L}^{w}_{Focal}$, $\mathcal{L}^{w}_{BCE}$, and $\mathcal{L}^{w}_{IoU}$ are defined in Eq. 3, Eq. 2, and Eq. 1, respectively. Using our proposed loss function helps ESFPNet reducing the importance of easy data samples and emphasizes learning from more challenging ones. This ensures the model focuses on learning diverse and complex regions. As a result, the accuracy of ESFPNet for lesion segmentation is increased.

[1] it equals to 1 if the pixel belongs to the lesion area and 0 if it belongs to the background
[2] it is usually set as a positive value (e.g., 2)

4.2 Lasso Regularization

Lasso regularization is introduced as a regularization term in loss functions to reduce over-fitting [25]. It is calculated as Eq. 5

$$L = \sum_{(x,y)} l(f(x, W), y) + \lambda \sum_{(\gamma \in \Gamma)} |\gamma|, \tag{5}$$

where x and y are the input and output of the training set, W denotes the trainable weights, λ controls the trade-off between the normal training loss l and Lasso regularization γ [3]. Eq. 5 is known as "sparse training" because the value of γ tends to zero during the training process. After sparse training, channels with near-zero factors can be pruned by removing all their incoming and outgoing connections and corresponding weights, resulting in efficient network structures. Therefore, we leverage the Lasso regularization as to mitigate over-fitting during the model training process.

4.3 Lesion Segmentation Model

Our proposed model is built based on the ESFPNet model, which is introduced in Section 3. The input image is initially preprocessed to remove redundant information of the endoscopic equipment on the image by threshing the black corners. These are black borders shown in Fig. 1. Subsequently, it is resized to a fixed dimension, e.g., $352x352x3$, before being fitted into the ESFPNet backbone. The backbone network is responsible for extracting both shallow (local) and deep (global) features from the input image.

The training process of the ESFPNet is illustrated in Fig. 1. For each input image, the ESFPNet model generates a prediction mask. The FBI loss function is computed using this prediction mask and the ground truth mask. This computation is executed for all input samples within the training batch size of the dataset. Additionally, the Lasso regularization term is added to the loss function to encourage model sparsity by constraining the weights of uninformative features to zero. The input of Lasso is the model's weights. The resulting loss value and the Lasso value are used optimization the weights of the ESFPNet network by the gradient back-propagation algorithm.

In the testing and predicting processes, each input image is preprocessed as described above. Then, it inputs to the ESFPNet model to get the prediction mask. The prediction mask presents the lesion and normal (background) area by the white pixels and the black pixels, respectively.

5 EXPERIMENTAL SETTING

5.1 Experimental Dataset

The experiments are conducted on two challenging lesion datasets, i.e., the Esophageal cancer and the Peptic ulcer dataset. These datasets comprise diverse images of lesions along with their respective ground truth annotations. To ensure a fair evaluation, the each dataset was randomly splitted into training, validation, and testing sets. We use the common ratio as $8:1:1$ for partitioning data [46] where 80% of the data allocated for training, 10% for validation, and 10% for testing. Here, we describe these datasets.

[3] The value of λ in our experiment is $1e-6$

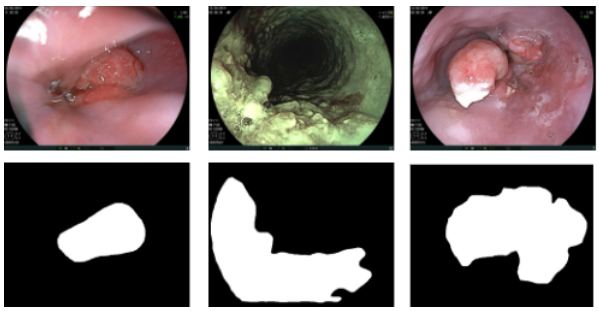

Figure 2: Esophageal cancer samples.

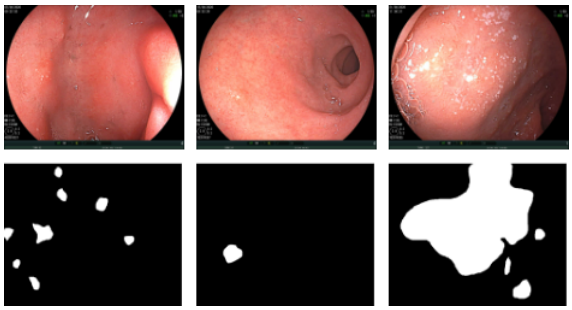

Figure 3: Peptic ulcer samples.

The Esophageal cancer and the Peptic ulcer dataset are two medical image datasets containing endoscopic images of the GI tract with diverse lesions and their corresponding ground truth for segmentation masks. These datasets were collected and labeled by experienced endoscopists. The Esophageal cancer dataset consists of 538 images with a size of 1280x995 pixels. Meanwhile, the Peptic ulcer dataset comprises 1159 images with a size of 1280x1024 pixels. Fig. 2 and Fig. 3 illustrate these two datasets.

As observed in these datasets, the size of lesions is both large and small, leading to imbalance between the lesion area and the normal area in endoscopic images, especially in the Peptic ulcer dataset. Additionally, the input images contain many noisy elements such as air bubbles, areas dazzled by light, or even regions obscured by endoscopic equipment. These factors significantly influence the performance of segmentation models.

5.2 Evaluation metrics

In this paper, we assess the performance of our proposed model using several commonly used evaluation metrics for image segmentation problems, namely Mean Intersection over Union (mIoU), Mean Dice coefficient (mDice), Mean Recall (mRecall), Mean Precision (mPrecision) [34]. Specifically, mDice calculates the balance between precision and recall for each class and reports the average across all classes. mIoU calculates the overlap between predicted and ground truth regions for each class, providing a class-specific evaluation. mRecall evaluates the model's ability to detect objects in different categories, while mPrecision measures its precision performance. These metrics offer a comprehensive assessment of the model's ability to accurately segment lesions in medical images.

5.3 Experimental Setups

We implement the model in PyTorch and accelerate training using NVIDIA GPUs. We trained these networks on an NVIDIA RTX 3060. We also employ random flipping, rotation, and brightness adjustments as data augmentation operations on the input image. The proposed loss function is used for training. Besides that, we also utilize the default AdamW optimizer [27] with a learning rate of 1×10^{-4} and trained our models for maximum as 200 epochs. The best model is selected based on the validation dice coefficient score on the validation dataset.

6 RESULT AND DISCUSSION

6.1 Accuracy

In this experiment, we aim to evaluate the ability of the lesion segmentation models on the collected datasets. To evaluate the proposed model, we conduct experiments for lesion segmentation on our collected datasets using the number of the state-of-the-art segmentation models, i.e., ResUnet++ [13], Resnet-Unet [16, 17], MSRFNet [41], FCBFormer [39]. The training sets, the validating sets, and the testing sets of these experiments are the same for all the lesion segmentaion models.

Table 1 presents the results of experimented models on the Esophageal cancer dataset. We can observe that the ESFPNet based models (ESPFNet [6] and our proposed model) achieve higher accuracy in almost metrics. This proves the effectiveness of ESFPNet architecture for the lesion segmentation problem. Moreover, our proposed model enhances the performance in terms of almost evaluation metrics. Specifically, the mDice and mIoU values of our proposed model are 0.8654 and 0.8332, respectively, indicating superior segmentation accuracy compared to other models. Besides, our method achieve highest mPrecision values, i.e., 0.9084. These results demonstrate that our proposed model enhance the accuracy of segmenting lesion areas in the endoscopy images.

Table 2 summarizes the performance of various models on the Peptic ulcer dataset. Evaluating in almost experimental metrics, the ESFPNet based models are able to segment the lesion area more effectively. This proves that the the ESFPNet architecture is effective for lesion segmentation on the Peptic ulcer dataset. Additionally, our proposed model outperforms the mDice, mRecall, mPrecision scores as 0.7005, 0.6973 and 0.7038, respectively, compared with the previous models. This signifies its outperformed segmentation capabilities in identifying the lesion regions. However, the mIoU value of our proposed model is slightly lower than that of the ESFPNet model as 0.003.

The proposed loss function helps to train the ESFPNet model more effectively. Moreover, using the new loss function only effects to the training process and does not effect to the inference time and the number of parameters of the model. This shows that our proposed model enhance the accuracy of the lesion segmentation model while remaining the model complexity compared to the original ESFPNet model.

6.2 Visualization

In the previous section, we demonstrate that our proposed model yields more accurate results on the two challenging GI lesion

Table 1: Results of all models on the Esophageal cancer dataset.

Metric	mDice	mIoU	mRecall	mPrecision
ResUNet++ [13]	0.7157	0.7018	0.7581	0.7495
ResNet34-UNet [17]	0.7204	0.5894	0.6667	0.8405
MSRFNet [41]	0.3216	0.2469	0.4248	0.3711
FCBFormer [38]	0.4285	0.3516	0.4592	0.4017
ESFPNet [6]	0.8633	0.8291	**0.8479**	0.8794
Ours	**0.8654**	**0.8332**	0.8262	**0.9084**

Table 2: Results of all models on the Peptic ulcer dataset.

Metric	mDice	mIoU	mRecall	mPrecision
ResUNet++ [13]	0.3912	0.5870	0.3812	0.5641
ResNet34-UNet [17]	0.4266	0.2992	0.4715	0.5436
MSRFNet [41]	0.1089	0.0815	0.1631	0.1402
FCBFormer [38]	0.2615	0.3901	0.1722	0.5433
ESFPNet [6]	0.6892	**0.7239**	0.6966	0.6820
Ours	**0.7005**	0.7236	**0.6973**	**0.7038**

datasets. In this section, we will visualize the segmentation results of experimental models for comparison.

Fig. 4 and Fig. 5 demonstrate that our proposed model helps predicting lesion regions of data samples more accurately. The reason of the less effective performance of the previous models is that they usually have complex architectures, leading to ineffective on our small dataset due to the over-fitting problem. To address this issue, we employ the regularization term, i.e., Lasso regularization. Furthermore, by incorporating our proposed loss function, the accuracy of the ESFPNet model is improved for the lesion segmentation problem, particularly for small-sized lesions.

7 CONCLUSION

In this research, we investigated the new loss function which aimed to enhance the performance of the lesion segmentation model. We introduced two key components to achieve this objective, i.e., the FBI loss and the Lasso regularization. The proposed loss function plays a crucial role in training the segmentation model for the lesions of varying sizes during the training phase. Moreover, the Lasso regularization helped to enable the sparse training, leading to reducing the over-fitting problem. Our experimental results proved the effectiveness of the proposed model for lesion segmentation on two collected datasets. However, the endoscopy images usually appears specular highlight making less effective of lesion segmentation models. In future work, we will apply the image processing techniques to remove the specular highlight pixels in the images to enhance the quality of the images before training.

REFERENCES

[1] Saruar Alam, Nikhil Kumar Tomar, Aarati Thakur, Debesh Jha, and Ashish Rauniyar. 2020. Automatic polyp segmentation using u-net-resnet50. *arXiv preprint arXiv:2012.15247* (2020).

[2] Md Zahangir Alom, Mahmudul Hasan, Chris Yakopcic, Tarek M Taha, and Vijayan K Asari. 2018. Recurrent residual convolutional neural network based on u-net (r2u-net) for medical image segmentation. *arXiv preprint arXiv:1802.06955* (2018).

[3] Jorge Bernal, Nima Tajkbaksh, Francisco Javier Sánchez, Bogdan J. Matuszewski, Hao Chen, Lequan Yu, Quentin Angermann, Olivier Romain, Bjørn Rustad, Ilangko Balasingham, Konstantin Pogorelov, Sungbin Choi, Quentin Debard, Lena Maier-Hein, Stefanie Speidel, Danail Stoyanov, Patrick Brandao, Henry Córdova, Cristina Sánchez-Montes, Suryakanth R. Gurudu, Gloria Fernández-Esparrach, Xavier Dray, Jianming Liang, and Aymeric Histace. 2017. Comparative Validation of Polyp Detection Methods in Video Colonoscopy: Results From the MICCAI 2015 Endoscopic Vision Challenge. *IEEE Transactions on Medical Imaging* 36, 6 (2017), 1231–1249. https://doi.org/10.1109/TMI.2017.2664042

[4] Leon Bungert, Tim Roith, Daniel Tenbrinck, and Martin Burger. 2022. A Bregman learning framework for sparse neural networks. *The Journal of Machine Learning Research* 23, 1 (2022), 8673–8715.

[5] F. Caliva, C. Iriondo, A.M. Martinez, S. Majumdar, and V. Pedoia. 2019. Distance map loss penalty term for semantic segmentation. In *Proceedings of The 2nd International Conference on Medical Imaging with Deep Learning*.

[6] Qi Chang, Danish Ahmad, Jennifer Toth, Rebecca Bascom, and William E Higgins. 2022. ESFPNet: efficient deep learning architecture for real-time lesion segmentation in autofluorescence bronchoscopic video. *arXiv preprint arXiv:2207.07759* (2022).

[7] Qi Chang, Danish Ahmad, Jennifer Toth, Rebecca Bascom, and William E Higgins. 2023. ESFPNet: efficient deep learning architecture for real-time lesion segmentation in autofluorescence bronchoscopic video. In *Medical Imaging 2023: Biomedical Applications in Molecular, Structural, and Functional Imaging*, Vol. 12468. SPIE, 1246803.

[8] Deng-Ping Fan, Ge-Peng Ji, Tao Zhou, Geng Chen, Huazhu Fu, Jianbing Shen, and Ling Shao. 2020. Pranet: Parallel reverse attention network for polyp segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part VI 23*. Springer, 263–273.

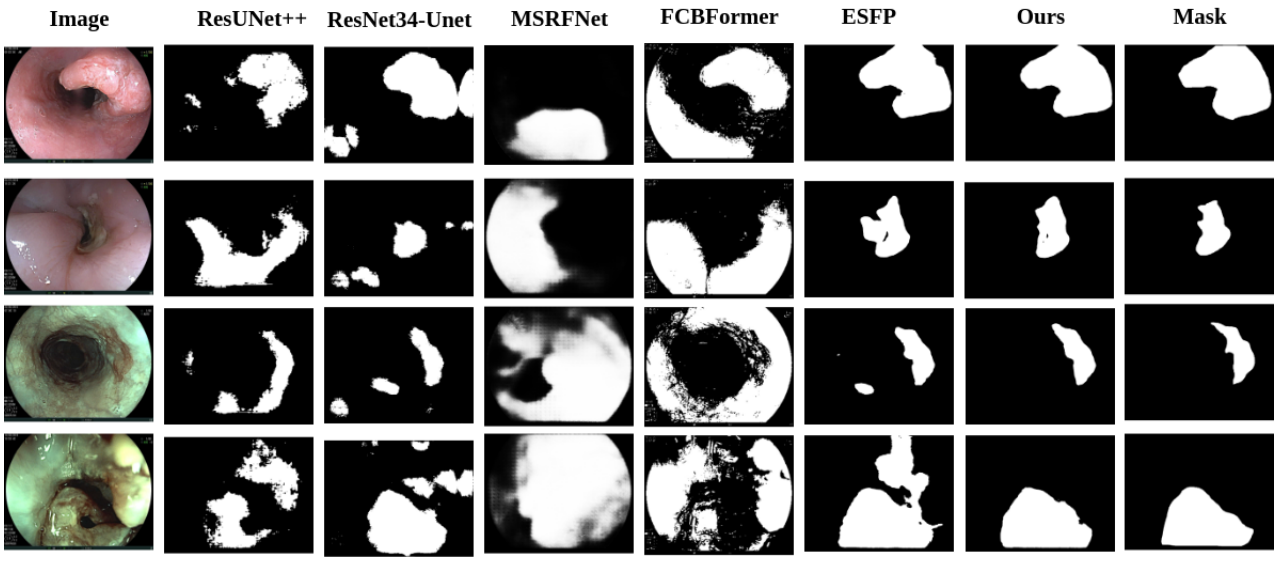

Figure 4: Visualization of Esophageal cancer predictions.

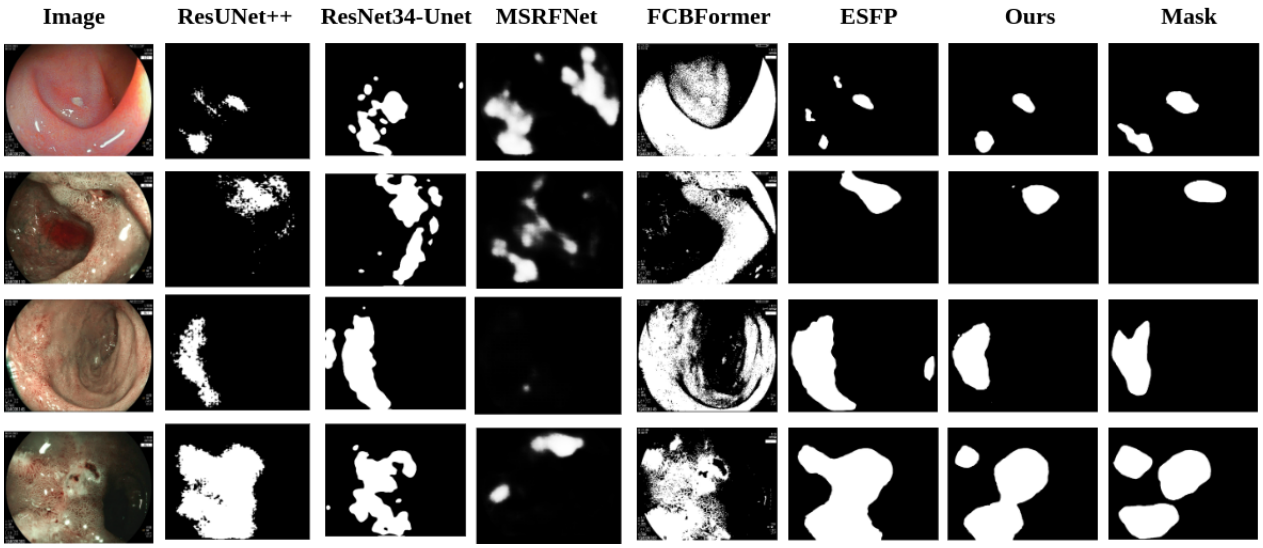

Figure 5: Visualization of Peptic ulcer predictions.

[9] Kerr Fitzgerald and Bogdan Matuszewski. 2023. FCB-SwinV2 Transformer for Polyp Segmentation. arXiv:2302.01027 [cs.CV]

[10] Elad Hazan and Tomer Koren. 2017. Proximal Algorithms for Learning Sparse Neural Networks. In *Proceedings of the 34th International Conference on Machine Learning*. 1496–1505.

[11] Hongliang He, Chi Zhang, Jie Chen, Ruizhe Geng, Luyang Chen, Yongsheng Liang, Yanchang Lu, Jihua Wu, and Yongjie Xu. 2021. A hybrid-attention nested UNet for nuclear segmentation in histopathological images. *Frontiers in Molecular Biosciences* 8 (2021), 614174.

[12] Debesh Jha, Michael A Riegler, Dag Johansen, Pål Halvorsen, and Håvard D Johansen. 2020. Doubleu-net: A deep convolutional neural network for medical image segmentation. In *2020 IEEE 33rd International symposium on computer-based medical systems (CBMS)*. IEEE, 558–564.

[13] Debesh Jha, Pia H Smedsrud, Michael A Riegler, Dag Johansen, Thomas De Lange, Pål Halvorsen, and Håvard D Johansen. 2019. Resunet++: An advanced architecture for medical image segmentation. In *2019 IEEE International Symposium on Multimedia (ISM)*. IEEE, 225–2255.

[14] Davood Karimi and Septimiu E. Salcudean. 2019. Reducing the Hausdorff Distance in Medical Image Segmentation with Convolutional Neural Networks. arXiv:1904.10030 [eess.IV]

[15] Hoel Kervadec, Jihene Bouchtiba, Christian Desrosiers, Eric Granger, Jose Dolz, and Ismail Ben Ayed. 2021. Boundary loss for highly unbalanced segmentation. *Medical Image Analysis* 67 (jan 2021), 101851. https://doi.org/10.1016/j.media.2020.101851

[16] Stephen L. H. Lau, Edwin K. P. Chong, Xu Yang, and Xin Wang. 2020. Automated Pavement Crack Segmentation Using U-Net-Based Convolutional Neural Network. *IEEE Access* 8 (2020), 114892–114899. https://doi.org/10.1109/access.2020.3003638

[17] Jean Le'Clerc Arrastia, Nick Heilenkötter, Daniel Otero Baguer, Lena Hauberg-Lotte, Tobias Boskamp, Sonja Hetzer, Nicole Duschner, Jörg Schaller, and Peter Maass. 2021. Deeply Supervised UNet for Semantic Segmentation to Assist Dermatopathological Assessment of Basal Cell Carcinoma. *Journal of Imaging* 7

(04 2021), 71. https://doi.org/10.3390/jimaging7040071
[18] Suck-Ho Lee, Il-Kwun Chung, Sun-Joo Kim, Jin-Oh Kim, Bong-Min Ko, Young Hwangbo, Won Ho Kim, Dong Hun Park, Sang Kil Lee, Cheol Hee Park, et al. 2008. An adequate level of training for technical competence in screening and diagnostic colonoscopy: a prospective multicenter evaluation of the learning curve. *Gastrointestinal endoscopy* 67, 4 (2008), 683–689.
[19] AM Leufkens, MGH Van Oijen, FP Vleggaar, and PD Siersema. 2012. Factors influencing the miss rate of polyps in a back-to-back colonoscopy study. *Endoscopy* 44, 05 (2012), 470–475.
[20] Mingchen Li, Xuechen Zhang, Christos Thrampoulidis, Jiasi Chen, and Samet Oymak. 2021. Autobalance: Optimized loss functions for imbalanced data. *Advances in Neural Information Processing Systems* 34 (2021), 3163–3177.
[21] Rui Li, Shunyi Zheng, Ce Zhang, Chenxi Duan, Jianlin Su, Libo Wang, and Peter M. Atkinson. 2022. Multiattention Network for Semantic Segmentation of Fine-Resolution Remote Sensing Images. *IEEE Transactions on Geoscience and Remote Sensing* 60 (2022), 1–13. https://doi.org/10.1109/tgrs.2021.3093977
[22] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal Loss for Dense Object Detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2980–2988. https://doi.org/10.1109/ICCV.2017.322
[23] Jie Liu, Weiyu Xu, and Ming Yan. 2015. Structured Sparsity through Convex Optimization: Fast Convergence using Active-Set Methods. In *Advances in Neural Information Processing Systems 28*. 3674–3682.
[24] Xinyu Liu and Yixuan Yuan. 2022. A Source-Free Domain Adaptive Polyp Detection Framework With Style Diversification Flow. *IEEE Transactions on Medical Imaging* 41, 7 (2022), 1897–1908. https://doi.org/10.1109/TMI.2022.3150435
[25] Zhuang Liu, Jianguo Li, Zhiqiang Shen, Gao Huang, Shouhai Yan, and Changshui Zhang. 2017. Learning Efficient Convolutional Networks through Network Slimming. In *Proceedings of the IEEE International Conference on Computer Vision*.
[26] Jonathan Long, Evan Shelhamer, and Trevor Darrell. 2015. Fully convolutional networks for semantic segmentation. 3431–3440. https://doi.org/10.1109/CVPR.2015.7298965
[27] Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. *arXiv preprint arXiv:1711.05101* (2019).
[28] Ange Lou, Shuyue Guan, Hanseok Ko, and Murray H. Loew. 2022. CaraNet: context axial reverse attention network for segmentation of small medical objects. In *Medical Imaging 2022: Image Processing*, Vol. 12032. International Society for Optics and Photonics, SPIE, 81 – 92. https://doi.org/10.1117/12.2611802
[29] Ange Lou, Shuyue Guan, and Murray Loew. 2023. CaraNet: context axial reverse attention network for segmentation of small medical objects. *Journal of Medical Imaging* 10, 1 (2023), 014005.
[30] Anran Lou and Murray Loew. 2021. Cfpnet: Channel-wise feature pyramid for real-time semantic segmentation. *arXiv preprint arXiv:2103.12212* (2021).
[31] Ange Lou and Murray Loew. 2021. CFPNET: Channel-Wise Feature Pyramid For Real-Time Semantic Segmentation. In *2021 IEEE International Conference on Image Processing (ICIP)*. 1894–1898. https://doi.org/10.1109/ICIP42928.2021.9506485
[32] Jun Ma, Jianan Chen, Matthew Ng, Rui Huang, Yu Li, Chen Li, Xiaoping Yang, and Anne L. Martel. 2021. Loss Odyssey in Medical Image Segmentation. *Medical Image Analysis* 71 (2021), 102035. https://doi.org/10.1016/j.media.2021.102035
[33] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. 2016. V-Net: Fully convolutional neural networks for volumetric medical image segmentation. *arXiv preprint arXiv:1606.04797* (2016).
[34] Dominik Müller, Iñaki Soto-Rey, and Frank Kramer. 2022. Towards a Guideline for Evaluation Metrics in Medical Image Segmentation. arXiv:2202.05273 [eess.IV]
[35] M.A. Rahman and Y. Wang. 2016. Optimizing intersection-over-union in deep neural networks for image segmentation. In *International symposium on visual computing*. 234–244.
[36] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. arXiv:1505.04597 [cs.CV]
[37] S.S.M. Salehi, D. Erdogmus, and A. Gholipour. 2019. Tversky loss function for image segmentation using 3d fully convolutional deep networks. In *International Workshop on Machine Learning in Medical Imaging*. 379–387.
[38] Edward Sanderson and Bogdan J. Matuszewski. 2022. FCN-Transformer Feature Fusion for Polyp Segmentation. In *Medical Image Understanding and Analysis*. Springer International Publishing, 892–907. https://doi.org/10.1007/978-3-031-12053-4_65
[39] Edward Sanderson and Bogdan J Matuszewski. 2022. FCN-transformer feature fusion for polyp segmentation. In *Medical Image Understanding and Analysis: 26th Annual Conference, MIUA 2022, Cambridge, UK, July 27–29, 2022, Proceedings*. Springer, 892–907.
[40] Jo Schlemper, Ozan Oktay, Michiel Schaap, Mattias Heinrich, Bernhard Kainz, Ben Glocker, and Daniel Rueckert. 2019. Attention gated networks: Learning to leverage salient regions in medical images. *Medical image analysis* 53 (2019), 197–207.
[41] Abhishek Srivastava, Debesh Jha, Sukalpa Chanda, Umapada Pal, Håvard D. Johansen, Dag Johansen, Michael A. Riegler, Sharib Ali, and Pål Halvorsen. 2022. MSRF-Net: A Multi-Scale Residual Fusion Network for Biomedical Image Segmentation. arXiv:2105.07451 [eess.IV]
[42] Y. Su, K. Jihoon, and K. Young-Hak. 2019. Major vessel segmentation on x-ray coronary angiography using deep networks with a novel penalty loss function. In *Proceedings of The 2nd International Conference on Medical Imaging with Deep Learning*.
[43] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2818–2826.
[44] Robert Tibshirani. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* 58, 1 (1996), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x
[45] Robert Tibshirani. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* 58, 1 (1996), 267–288.
[46] Fang Wang and Weibin Hong. 2022. Polyp DataSet. https://doi.org/10.6084/m9.figshare.21221579.v2.
[47] Z. Wu, C. Shen, and A. van den Hengel. 2016. Bridging category-level and instance-level semantic image segmentation. *arXiv preprint arXiv:1605.06885* (2016).
[48] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. 2021. Segformer: Simple and efficient design for semantic segmentation with transformers. In *Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 34. Curran Associates, Inc., 12077–12090.
[49] Li Yuan, Yuxuan Chen, Tao Wang, Weihao Yu, Yihong Shi, Zheng-Hua Jiang, Francis E. Tay, Jiashi Feng, and Shuicheng Yan. 2021. Tokens-to-token ViT: Training vision transformers from scratch on ImageNet. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 558–567.
[50] Dongdong Zhao, Weihao Ge, Peng Chen, Yingtian Hu, Yuanjie Dang, Ronghua Liang, and Xinxin Guo. 2022. Feature Pyramid U-Net with Attention for Semantic Segmentation of Forward-Looking Sonar Images. *Sensors* 22, 21 (2022). https://doi.org/10.3390/s22218468
[51] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. 2019. Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. *IEEE transactions on medical imaging* 39, 6 (2019), 1856–1867.