International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

ISSN:2147-6799 www.ijisae.org Original Research Paper

# MSDFN (Multi – Scale Dilated Fusion Network) for Automatic Instrument Segmentation

**Wangkheirakpam Reema Devi[a,*], Sudipta Roy[a], Khelchandra Thongam[b] , Chiranjiv Chingangbam[c]**

**Submitted:** 10/09/2022 **Accepted:** 20/12/2022

**Abstract**

With the recent advancements in the field of semantic segmentation, an encoderdecoder approach like U-Net are most widely used to solve biomedical image segmentation tasks. To improve upon the existing U-Net, we proposed a novel architecture called Multi-Scale Dilated Fusion Network (MSDFNet). In this work, we have used the pre-trained ResNet50 as the encoder, which had already learned features that can be used by the decoder to generate the binary mask. In addition, we had used skip-connections to facilitate the transfer of features from the encoder to the decoder directly. Some of these features are lost due to the depth of the network. The decoder consists of a Multi-Scale Dilated Fusion block, as the main components of the decoder, where we fused the multi-scale features and then apply some dilated convolution upon them. We have trained both the UNet and the proposed architecture on the Ksavir-Instrument dataset, where the proposed architecture has a 3.701 % gain in the F1 score and 4.376 % in the Jaccard. These results show the improvement over the existing U-Net model.

**Keywords**: UNet, MSDFNet, medical image segmentation, Deep learning, multiscale dilated convolution

## 1. Introduction

Image segmentation is the most widely used area of the computer vision research community. It is the process of identifying the class label for each pixel of the image and determining the region of interest. Image segmentation helps

to determine the areas of interest within an image, frame or even a video with pixel level accuracy. Due to its pixel-level accuracy, image segmentation has gained the attention of researchers in the medical domain. The researchers used it to diagnose various diseases, identify medical conditions and even use it for the identification of a wide range of surgical instruments and parts during a surgical procedure.

Robotic-assisted surgery (RAS) [1] has evolved significantly over the years and has improved surgical performance along with patient safety. Instrument segmentation helps in better identification of the surgical instrument along with various other tools used during a surgical procedure. This helps in providing optimal instrument control and minimizes unnecessary risks. Although, there also exist some factors which hinder the optimal performance of the procedure, such as the artefacts caused by the motion, specularities , debris, low contrast, bubbles, bodily fluids and blood. As a result, clinical endoscopist face hurdles in visual interpretation. So it becomes necessary to develop a method that

can solve the above-mentioned constraints. To encounter the above issues we explore the U-Net architecture which is most widely used for medical image segmentation. This architecture has performed well but in some cases, it remains a barrier to achieving accurate segmentation. We overcome the obstacle and develop a new architecture called Multi-Scale Dilated Fusion Network (MSDFNet) that can efficiently achieve accurate segmentation performance.

The main contributions of our paper can be summarized as the following:

1. We proposed a novel architecture called Multi-Scale Dilated Fusion Network (MSDFNet) which is an encoder-decoder based segmentation architecture mainly built using a pre-trained encoder and fusion of multi-scale features. Moreover dilated convolutions are also used in the decoder network to further improve the performance.
2. We evaluate the proposed MSDFNet on the Kvasir-instrument [2] dataset for instrument segmentation, where it shows a significant improvement over the existing U-Net.

[a]*Assam University*
[b]*NIT Manipur* , [c]*Manipur Technical University*
[*]*Corresponding author*
*Email addresses: reemawng21@gmail.com (Wangkheirakpam Reema Devi)*

## 2. Related work

For the instrument segmentation , a variety of traditional techniques have been presented, ranging from simple background subtraction and colour threshold procedures to far more advanced approaches, such as marked point processes. Many of these techniques have lately been re-examined .

Convolutional neural networks have led to the development of alternative approaches. Deep neural networks have become the state-of-the-art model for object recognition thanks to recent breakthroughs in deep neural networks, particularly in their optimization. Deep neural networks have also been employed for semantic segmentation, with employing "de-convolution layers" and upsampling to recognise and precisely pinpoint items within a photograph. Different architectures derived from various intuitions are also possible and have been used in this paper.

| **Dataset** | **Content** | **Task type** | **Procedure** |
|---|---|---|---|
| Instrument segmentation and tracking (2015) [6] | Rigid and robotic instruments [6] | Segmentation and tracking [6] | Laparoscopy [6] |
| Robotic Instrument Segmentation (2017) [15] | Robotic surgical instruments [15] | Binary segmentation, part based segmentation, instrument segmentation [15] | Abdominal porcine [15] |
| Robotic Scene Segmentation (2018) [16] | Surgical instruments and other [16] | Multi-instance segmentation [16] | Robotic nephrectomy [16] |
| Robust Medical instrument segmentation (2019) [17] | laparoscopic instrument [17] | Binary segmentation, multiple instance detection, multiple instance segmentation[17] | Laparoscopy [17] |
| Kvasir-Instrument | Diagnostic and therapeutic tools in endoscopic images | Binary segmentation Detection and localization | Gastroscopy & colonoscopy |

**Table 1**: Available instrument datasets

Instrument segmentation [3] is an important area of research in the field of surgical technology and, especially, in the field of robotic-assisted surgery. In the context of instrument segmentation and tracking, the Endoscopic vision (EndoVis) challenge have greatly contributed by providing a labelled dataset for the research. The challenge aims at developing novel approaches to improve the performance and efficiency of robotic surgery, particularly instrument segmentation. With the exception of 2016, the Medical Image Computing and Computer Assisted Intervention Society (MICCAI) has hosted an endoscopic vision (EndoVis) challenge since 2015. There are several sub-challenges in the Endovis challenge. The challenge has year-byyear information regarding the hosted sub-challenge. Bodenstedt et al. [14] established the "EndoVis 2015 Instrument sub-challenge" to test novel methodologies and assess the Machine Learning (ML) algorithm for instrument segmentation and tracking on a common dataset. The organisers were given two duties to complete: (1) segmentation and (2) tracking. The purpose of the challenge was to provide a solution to the problem of segmenting and monitoring articulated devices in both laparoscopic and robotic surgery. A thorough examination of the methodologies utilised in instrument segmentation and tracking for minimally invasive procedures. This challenge is held as an international event starting in the year 2015 at the Medical Image Computing and Computer-Assisted Intervention Society (MICCAI). With the exception of 2016, the challenges were held in 2017, 2018 and 2019. Deep learning performs effectively for instrument segmentation and tracking tasks, according to the extensive testing. For more information about the Endoscopic vision (EndoVis) challenge, please visit https://endovis.grandchallenge.org.

Most of the participants used U-Net[6] and ResNet[7]. The author[https://arxiv.org/pdf/2107.02319.pdf] has published a detailed research paper on instrument segmentation using the ROBUST-MIS challenge 2019 dataset. The paper used the U-Net, RIC-Unet [8], FCN[9], [10], DeconvNet [11]to analyze the performance of all the methods. So, a more accurate medical image segmentation approach is needed . In order to overcome this need, we have proposed MSDFNet architecture and compared with various state-ofthe-art architecture. Our proposed model produces significant output segmentation masks despite the challenging images.

## 3. Method

In this section, we are going to present the proposed Multi-Scale Dilated Fusion Network (MSDFNet) architecture in detail. We present the main components required to build the architecture.

*Residual block:*. As we increase the depth of the neural network, the performance also increases. The increase in performance is up to a certain limit, after that, the performance starts to decrease. This decrease in performance is due to the vanishing or exploding gradients problem. The problem is solved with the use of residual blocks [13], which introduces a shortcut connection between the layers, thus avoiding the gradient problem. The residual block consists of two

3x3 convolution blocks and a shortcut connection (identity mapping). Each convolution block begins a 3x3 convolution layer, followed by batch normalization and a Rectified Linear Unit (ReLU) activation function.

*Multi-scale dilated fusion (MSDF) block:*. The proposed MSDF block is used in the decoder part of the network. The block begins with the fusion of the multiscale features into a single multi-scale feature. Here two feature maps of different spatial resolutions are taken from the pre-trained ResNet50[cite] encoder and a bilinear upsampling is performed to get new feature maps of similar size. These features are then followed by the residual block and then concatenated along with the third feature from the encoder. Next, the concatenated feature map is followed by 1x1 convolution, along with the batch normalization and a ReLU activation function. After that, two parallel convolution layers with a dilation rate of 3 and 9 are applied followed by a concatenation. Finally, we have a 1x1 convolution, followed by batch normalization and a ReLU activation function.

*Proposed MSDFNet architecture:*. The proposed architecture is illustrated in Figure 1 presenting all the main components. It uses an encoder-decoder design

which is commonly used by most of the image segmentation architecture. In this

architecture, we combine the strength of the pre-trained ResNet50, residual block and the MSDF block which allows for better performance when compared with the baseline architecture. The proposed architecture is a fully convolutional network consisting of a pre-trained ResNet50 encoder and decoder. The network begins with a pre-trained ResNet50 as an encoder, which is trained on the ImageNet dataset. The encoder takes the input image encodes it into a compressed representation using multiple convolution layers and pooling layers.

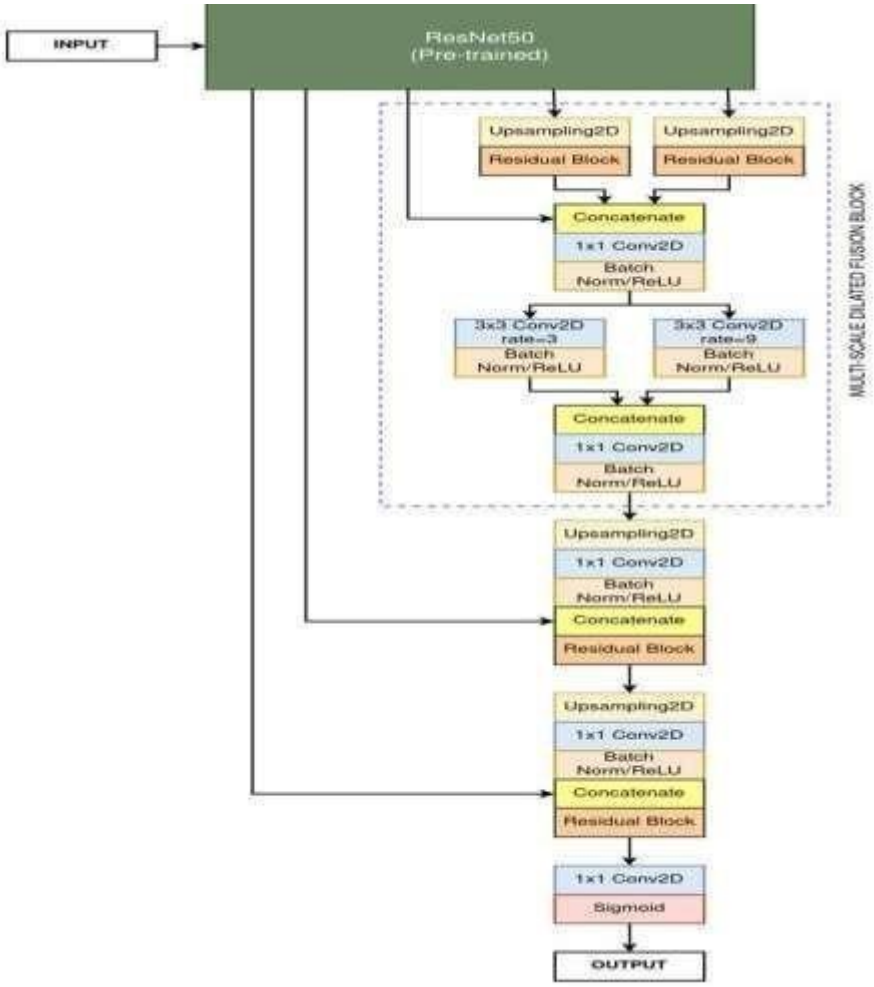


Figure 1: The block diagram of the MSDFNet Architecture

This compressed representation of the input image is passed to the MSDF block, where we combine the features from different scales to create a better representation. This representation is further enhanced by the dilated convolution, where we increase the field of view to grasp better information. The resulting feature map is then upsampled by using the bilinear interpolation and then concatenated the skip connection from the encoder. These skip connection helps to get the low-level features from the encoder to the decoder. After skip connection comes the residual block for learning better representation. At last, a 1x1 convolution is used along with the sigmoid activation function to get the final segmentation mask.

## 4. Dataset and Evaluation Metrics

To evaluate our proposed MSDFNet, we have chosen the Kvasir-instrument[2] dataset. The dataset consists of a variety of instrument images taken from different colonoscopy procedures. It includes images of different gastrointestinal tools such as snares, balloons and biopsy forceps, and many more. The dataset contains 590 images and their respective annotated masks, which were verified by two expert gastrointestinal endoscopists. The performance of the baseline U-Net and the newly proposed MSDFNet are compared using the standard image segmentation metrics, such as F1 (Dice Coefficient), mean Intersection over Union (mIoU). Additionally, we have also calculated Recall and Precision for each method.

## 5. Experimental setup

The baseline U-Net and the proposed MSDFNet architecture were trained in the same environment using a similar configuration. Both the architectures are implemented using TensorFlow 2.5 framework in python 3.8. For a fair comparison, we have split the Kvasir-Instrument dataset into training, validation and testing datasets in the ratio of 60:20:20. For the training, we have 354 images and masks pairs, while the validation and testing dataset contains 118 images and masks pairs, respectively. All the images are resized to a resolution of 256 x 256 pixels. The models are trained using a GTX 1040 Ti 4GB GPU with a batch size of four. The dice-coefficient loss is used to calculate the error rate,

which is optimized by the Adam optimizer along with a learning rate of 1e-4.

## 6. Experimental Results

In this section, we are going to present the results of the baseline architecture U-Net and the newly proposed MSDFNet architecture and see the improvements made. The quantitative results on the Kvasir-Instrument dataset are presented in Table 2. From the table1, we can see that the overall performance of the proposed architecture is significantly better than the U-Net architecture.

| Method | F1 | Jaccard | Precision | Recall | Accuracy |
|---|---|---|---|---|---|
| U-Net | 85.64 | 84.73 | 89.05 | 92.85 | 96.27 |
| RIC-Unet | 92.78 | 86.35 | 91.02 | 92.89 | 96.36 |
| ResNet | 89.64 | 85.72 | 85.81 | 85.67 | 96.80 |
| DeconvNet | 86.60 | 85.70 | 87.80 | 85.50 | 95.80 |
| **MSDFNet** | **93.34** | **89.10** | **92.20** | **95.32** | **98.76** |

**Table 1**: Comparision between UNet and various state-of-the-art architecture

The proposed MSDFNet has achieved an F1 score of 93.34, mIoU of 89.10 and a pixel accuracy of 98.76. When compared with the U-Net architecture, the proposed MSDFNet has achieved a 3.7% higher F1 score and 4.3% higher mIoU. Additionally, we have also calculated recall and precision which are 92.20 and 95.32 respectively. Figure 3 shows the qualitative results, where we can see that the proposed MSDFNet has better masks and more accurate edge detection performance.

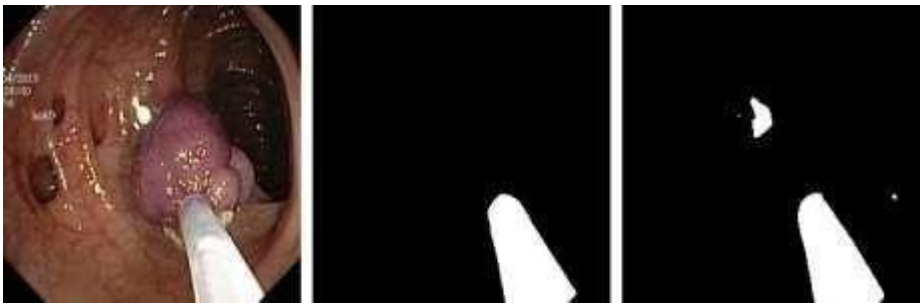

**Figure 2**: UNet result

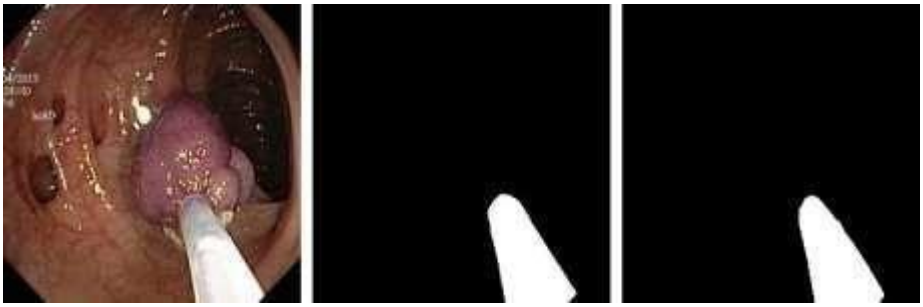

**Figure 3**: MSDFNet result

## 7. Discussion

The proposed MSDFNet produce acceptable results on the Kvasir-Instrument dataset. Figure 3 proves that the segmentation masks predicted by the MSDFNet are superior to the baseline U-Net architecture in terms of details, with better quality. We trained the proposed model with different hyperparameters, including different loss functions, optimizers, and many more. Finally, after a set of experiments, we found the best configuration of the hyperparameters that gives the best performance. However, we also observed that the batch size, loss function, number of filers and dilation rate can influence the results of the MSDFNet. So, we can say that the performance of the proposed model can be done by adjusting the architecture and searching for the best set of hyper parameters. We conclude the application of the MSDFNet has the potential to solve numerous segmentation tasks outside the medical domain. It can be used in natural image segmentation and other image segmentation tasks.

## 8. Conclusion

In this paper, we present the MSDFNet, an architecture for the gastro-in testinal instrument segmentation task. The architecture address the need for accurate segmentation of instruments. The architecture takes the advantage of the pre-trained encoder, multi-scale function and dilated convolution to tackle segmentation problems. In the future, we plan to experiment with more pretrained models and also try to take more advantage of the un-labeled data. As the unlabeled data is easily available and we don't need to spend time and resources to annotate it. We can try to add some attention mechanisms to further improve the performance of the model.

## 9. Conflict of Interest

The authors have no conflicts of interest to declare that they are relevant to the content of this article.

## References


[1]. J. D. Wright, Robotic-assisted surgery: balancing evidence and implementation, Jama 318 (16) (2017) 1545–1547.

[2]. D. Jha, S. Ali, K. Emanuelsen, S. A. Hicks, V. Thambawita, E. Garcia-Ceja, M. A. Riegler, T. de Lange, P. T. Schmidt, H. D. Johansen, et al., Kvasirinstrument: Diagnostic and therapeutic tool segmentation dataset in gastrointestinal endoscopy, in: International Conference on Multimedia Modeling, Springer, 2021, pp. 218–229.

[3]. Kumari, S. S. ., and K. S. . Rani. "Big Data Classification of Ultrasound Doppler Scan Images Using a Decision Tree Classifier Based on Maximally Stable Region Feature Points". International Journal on Recent and Innovation Trends in Computing and Communication, vol. 10, no. 8, Aug. 2022, pp. 76-87, doi:10.17762/ijritcc.v10i8.5679.

[4]. A. A. Shvets, A. Rakhlin, A. A. Kalinin, V. I. Iglovikov, Automatic instru-ment segmentation in robot-assisted surgery using deep learning, in: 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA), IEEE, 2018, pp. 624–628.

[5]. K. He, G. Gkioxari, P. Doll´ar, R. Girshick, Mask r-cnn, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 2961–2969.

[6]. V. Balasubramanian, R. Kumar, S. J. Kamireddi, R. Sathish, D. Sheet, Semantic segmentation, detection and localisation of mucosal lesions from gastrointestinal endoscopic images using sumnet., in: EndoCV@ ISBI, 2020, pp. 82–83.

[7]. O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical image computing and computer-assisted intervention, Springer, 2015, pp. 234–241.

[8]. Ghazaly, N. M. . (2022). Data Catalogue Approaches, Implementation and Adoption: A Study of Purpose of Data Catalogue. International Journal on Future Revolution in

Computer Science &Amp; Communication Engineering, 8(1), 01–04. https://doi.org/10.17762/ijfrcsce.v8i1.2063

[9]. A. C¸inar, M. Yildirim, Detection of tumors on brain mri images using the 190 hybrid convolutional neural network architecture, Medical hypotheses 139 (2020) 109684.

[10]. X. Xiao, S. Lian, Z. Luo, S. Li, Weighted res-unet for high-quality retina vessel segmentation, in: 2018 9th international conference on information technology in medicine and education (ITME), IEEE, 2018, pp. 327–331.

[11]. Z. Cai, N. Vasconcelos, Cascade r-cnn: Delving into high quality object detection, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 6154–6162.

[12]. H. Oh, M. Lee, H. Kim, J. Paik, Metadata extraction using deeplab v3 and probabilistic latent semantic analysis for intelligent visual surveillance systems, in: 2020 IEEE International Conference on Consumer Electronics (ICCE), IEEE, 2020, pp. 1–2.

[13]. Kabisha, M. S., Rahim, K. A., Khaliluzzaman, M., & Khan, S. I. (2022). Face and Hand Gesture Recognition Based Person Identification System using Convolutional Neural Network. International Journal of Intelligent Systems and Applications in Engineering, 10(1), 105–115. https://doi.org/10.18201/ijisae.2022.273

[14]. J. Tang, J. Li, X. Xu, Segnet-based gland segmentation from colon cancer histology images, in: 2018 33rd Youth Academic Annual Conference of Chinese Association of Automation (YAC), IEEE, 2018, pp. 1078–1082.

[15]. N. K. Tomar, D. Jha, S. Ali, H. D. Johansen, D. Johansen, M. A. Riegler, P. Halvorsen, Ddanet: Dual decoder attention network for automatic polyp segmentation, arXiv preprint arXiv:2012.15245.

[16]. E. Goceri, Analysis of deep networks with residual blocks and different activation functions: classification of skin diseases, in: 2019 Ninth in-ternational conference on image processing theory, tools and applications (IPTA), IEEE, 2019, pp. 1–6.

[17]. M. J. Traum, J. Fiorentine. (2021). Rapid Evaluation On-Line Assessment of Student Learning Gains for Just-In-Time Course Modification. Journal of Online Engineering Education, 12(1), 06–13. Retrieved from http://onlineengineeringeducation.com/index.php/joee/article/view/45

[18]. Bodenstedt, S., Allan, M., Agustinos, A., Du, X., Garcia-Peraza-Herrera, L., Kenngott, H., Kurmann, T., M¨uller-Stich, B., Ourselin, S., Pakhomov, D., et al.: Comparative evaluation of instrument segmentation and tracking methods in minimally invasive surgery. arXiv preprint arXiv:1805.02475 (2018)

[19]. Allan, M., et al.: 2017 robotic instrument segmentation challenge. arXiv preprint arXiv:1902.06426 (2019)

[20]. Allan, M., Azizian, M.: Robotic scene segmentation sub-challenge. arXiv preprint arXiv:1902.06426 (2019)

[21]. Ross, T., et al.: Robust medical instrument segmentation challenge 2019. arXiv preprint arXiv:2003.10299 (2020)